About the Authors

Ann E. Watkins is Professor of Mathematics at California State University, Northridge (CSUN). She received her Ph.D. in education from the University of California, Los Angeles. She is a former president of the Mathematical Association of America (MAA) and a Fellow of the American Statistical Association (ASA). Dr. Watkins has served as co-editor of College Mathematics Journal, as a member of the Board of Editors of American Mathematical Monthly, and as Chair of the Advanced Placement Statistics Development Committee. She was selected as the 1994–1995 CSUN Outstanding Professor and won the 1997 CSUN Award for the Advancement of Teaching Effectiveness. Before moving to CSUN in 1990, she taught for the Los Angeles Unified School District and at Pierce College in Los Angeles. In addition to numerous journal articles, she is the co-author of books based on work produced by the Activity-Based Statistics Project (co-authored with Richard Scheaffer), the Quantitative Literacy Project, and the Core-Plus Mathematics Project.

Richard L. Scheaffer is Professor Emeritus of Statistics at the University of Florida, where he served as chairman of the Department of Statistics for 12 years. He received his Ph.D. in statistics from Florida State University. Dr. Scheaffer's research interests are in the areas of sampling and applied probability, especially in their applications to industrial processes. He has published numerous papers and is co-author of four college-level textbooks. In recent years, much of his effort has been directed toward statistics education at the elementary, secondary, and college levels. He was one of the developers of the Quantitative Literacy Project where he helped form the basis of the data analysis emphasis in mathematics curriculum standards recommended by the National Council of Teachers of Mathematics. Dr. Scheaffer has also directed the task force that developed the Advanced Placement Statistics Program and served as its first Chief Faculty Consultant. Dr. Scheaffer is Fellow and past president of the American Statistical Association, from which he received a Founders Award.

George W. Cobb is the Robert L. Rooke Professor of Statistics at Mt. Holyoke College, where he served a three-year term as Dean of Studies. He received his Ph.D. in statistics from Harvard University, and is an expert in statistics education with a significant publication record. He chaired the joint committee on undergraduate statistics of the Mathematical Association of America and the American Statistical Association and is a Fellow of the American Statistical Association. He also led the Statistical Thinking and Teaching Statistics (STATS) project of the Mathematical Association of America, which helped professors of mathematics learn to teach statistics. Over the past two decades, Dr. Cobb has frequently served as an expert witness in lawsuits involving alleged employment discrimination.
Acknowledgments

This book is a product of what we have learned from the statisticians and teachers who have been actively involved in helping the introductory statistics course evolve into one that emphasizes activity-based learning of statistical concepts while reflecting modern statistical practice. This book is written in the spirit of the recommendations from the MAA’s STATS project, the ASA’s Quantitative Literacy and GAISE projects, and the College Board’s AP Statistics course. We hope that it adequately reflects the wisdom and experience of those with whom we have worked and who have inspired and taught us.

We owe special thanks to Corey Andreasen, an outstanding high school mathematics and AP Statistics teacher at North High School in Sheboygan, Wisconsin, for his insight into what makes a topic “teachable” to high school students. His careful review of the manuscript led to many clarifications of wording and improvements in exercises, all of which will make it easier for you to learn the material. Corey also has made substantial contributions to the solutions and teacher’s notes, adding his unique perspective and sense of humor.

It has been an awesome experience to work with the Key Curriculum staff and field-test teachers, who always put the interests of students and teachers first. Their commitment to excellence has motivated us to do better than we ever could have done on our own. Steve, Casey, Jim, Kristin B., Kristin F., and the rest of the staff have been professional and astute throughout. Our deepest gratitude goes to Cindy Clements and Josephine Noah, the editors of the first and second editions, respectively, who have been a joy to work with. (Not all authors say that—and mean it—about their editors.) Cindy and Josephine were outstanding high school teachers before coming to Key. Their organizational skills, experience in the classroom, and insight have improved every chapter of this text.
<table>
<thead>
<tr>
<th>CHAPTER 5</th>
<th>Probability Models</th>
<th>286</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Constructing Models of Random Behavior</td>
<td>288</td>
</tr>
<tr>
<td>5.2</td>
<td>Using Simulation to Estimate Probabilities</td>
<td>301</td>
</tr>
<tr>
<td>5.3</td>
<td>The Addition Rule and Disjoint Events</td>
<td>315</td>
</tr>
<tr>
<td>5.4</td>
<td>Conditional Probability</td>
<td>325</td>
</tr>
<tr>
<td>5.5</td>
<td>Independent Events</td>
<td>339</td>
</tr>
<tr>
<td></td>
<td>Chapter Summary</td>
<td>349</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 6</th>
<th>Probability Distributions</th>
<th>356</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>Random Variables and Expected Value</td>
<td>358</td>
</tr>
<tr>
<td>6.2</td>
<td>The Binomial Distribution</td>
<td>382</td>
</tr>
<tr>
<td>6.3</td>
<td>The Geometric Distribution</td>
<td>393</td>
</tr>
<tr>
<td></td>
<td>Chapter Summary</td>
<td>402</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 7</th>
<th>Sampling Distributions</th>
<th>408</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1</td>
<td>Generating Sampling Distributions</td>
<td>410</td>
</tr>
<tr>
<td>7.2</td>
<td>Sampling Distribution of the Sample Mean</td>
<td>426</td>
</tr>
<tr>
<td>7.3</td>
<td>Sampling Distribution of the Sample Proportion</td>
<td>446</td>
</tr>
<tr>
<td></td>
<td>Chapter Summary</td>
<td>458</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 8</th>
<th>Inference for Proportions</th>
<th>466</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.1</td>
<td>Estimating a Proportion with Confidence</td>
<td>468</td>
</tr>
<tr>
<td>8.2</td>
<td>Testing a Proportion</td>
<td>490</td>
</tr>
<tr>
<td>8.3</td>
<td>A Confidence Interval for the Difference of Two Proportions</td>
<td>518</td>
</tr>
<tr>
<td>8.4</td>
<td>A Significance Test for the Difference of Two Proportions</td>
<td>526</td>
</tr>
<tr>
<td>8.5</td>
<td>Inference for Experiments</td>
<td>539</td>
</tr>
<tr>
<td></td>
<td>Chapter Summary</td>
<td>554</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 9</th>
<th>Inference for Means</th>
<th>560</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.1</td>
<td>A Confidence Interval for a Mean</td>
<td>562</td>
</tr>
<tr>
<td>9.2</td>
<td>A Significance Test for a Mean</td>
<td>580</td>
</tr>
<tr>
<td>9.3</td>
<td>When Things Aren’t Normal</td>
<td>602</td>
</tr>
<tr>
<td>9.4</td>
<td>Inference for the Difference Between Two Means</td>
<td>616</td>
</tr>
<tr>
<td>9.5</td>
<td>Paired Comparisons</td>
<td>640</td>
</tr>
<tr>
<td></td>
<td>Chapter Summary</td>
<td>663</td>
</tr>
</tbody>
</table>
Data enter the conversation whether you talk about income, sports, health, politics, the weather, or prices of goods and services. In fact, in this age of information technology, data come at you at such a rapid rate that you can catch only a glimpse of the masses of numbers. You cannot cope intelligently in this quantitative world unless you have an understanding of the basic concepts of statistics and have had practice making informed decisions using real data.

Statistics in Action is designed for students taking an introductory high school statistics course and includes all of the topics in the Advanced Placement (AP) Statistics syllabus. Beginning in Chapter 1 with a court case about age discrimination, you will be immersed in real-world problems that can be solved only with statistical methods. You will learn to explore, summarize, and display data; design surveys and experiments; use probability to understand random behavior; make inferences about populations by looking at samples from those populations; and make inferences about the effect of treatments from designed experiments.

After completing your statistics course, you will be prepared to take the AP Statistics Exam, to take a follow-up college-level course, and, above all, to make informed decisions in this world of data.

You will be using this book in a first course in statistics, so, you aren't required to know anything yet about statistics. You may find that your success in statistics results more from your perseverance in trying to understand what you read rather than your skill with algebra. However, basic topics from algebra, such as slope, linear equations, exponential equations, and the idea of a logarithm, will arise throughout the book. Be prepared to review these as you go along.

Statistics from a Modern Perspective

Statistical work is more interactive than it was a generation ago. Computers and graphing calculators have automated the graphical exploration of data and, in the process, have made the practice of statistics a more visual enterprise. Statistical techniques are also changing as simulations allow statisticians (and you) to shift the emphasis from following recipes for calculations to paying more attention to statistical concepts. Your instructor has selected this book so that you may learn this modern, data-analytic approach to statistics and because he or she encourages you to be an active participant in the classroom, wants you to see real data (if you have only pretend data, you can only pretend to analyze it), believes that statistical analyses must be tailored to the data, and uses graphing calculators or statistical software for data analysis and for simulations.
These features grow out of the vigorous changes that have been reshaping the practice of statistics and the teaching of statistics over the last quarter century.

The most basic question to ask about any data set is, “Where did the data come from?” Good data for statistical analysis must come from a good plan for data collection. Thus, Statistics in Action treats the design and analysis of experiments honestly and thoroughly and discusses how these methods of collecting data differ from observational studies. It then follows through on this theme by relating the statistical analysis to the manner in which the data were collected.

**What You Should Know About This Book**

Throughout Statistics in Action you will see many graphical displays, lots of real data, activities that introduce each major topic, computer printouts, questions for you to discuss with your class, and practice problems so that you can be sure you understand the basics before you move on. The practice problems are found at the end of every section, organized by topic. You should work every problem for those topics that you wish to learn. The answers to all practice problems and odd-numbered exercises are in the back of the book.

Also in the back of the book you will find a glossary of statistical notation, a glossary of statistical terminology, statistical tables, and an index so that you can locate topics quickly. Two tables are reprinted inside the back cover of the book.

You should keep in mind that the emphasis throughout is on the development of statistical concepts. Even when looking for a numerical answer in a practice problem or exercise, think about the underlying concept that is being illustrated. Concepts are carefully developed in the written text, so read the material thoughtfully and strive to fully understand what is being explained. Concepts are also developed in the discussion questions. These are designed for in-class discussions, but regard them as part of your reading assignment, even if class time is too limited for full discussion. Through reading, discussing, and working practice problems and exercises you will develop a profound understanding of statistical thinking, and that will serve you well as a basis for a lifetime of coping with a quantitative world.

*Ann Watkins*
*Dick Scheaffer*
*George Cobb*
Statistics in Action
Understanding a World of Data
Second Edition
Chapter 1

Statistical Reasoning: Investigating a Claim of Discrimination

Were older workers discriminated against during a company's downsizing? When an older worker felt he had been unfairly laid off, his lawyers called on a statistician to help them evaluate the claim.
In the year Robert Martin turned 54, the Westvaco Corporation, which makes paper products, decided to downsize. They laid off several members of the engineering department, including Robert Martin. Later that year, he sued Westvaco, claiming he had been laid off because of his age. A major piece of Martin’s case was based on a statistical analysis of the ages of the Westvaco employees.

In the two sections of this chapter, you will get a chance to try your hand at two very different kinds of statistical work, exploration and inference. Exploration is an informal, open-ended examination of data. Your goal in the first section will be to uncover and summarize patterns in data from Westvaco that bear on the Martin case. You will try to formulate and answer basic questions such as “Were those who were laid off older on average than those who weren’t laid off?” You can use any tools—graphs, averages, and so on—that you think might be useful. Inference, which you’ll use in the second section, is quite different from exploration in that it follows strict rules and focuses on judging whether the patterns you found are the sort you would expect. You’ll use inference to decide whether the patterns you find in the Westvaco data are the sort you would expect from a company that does not discriminate on the basis of age, or whether further investigation into possible age discrimination is needed.

The purpose of this first chapter is to familiarize you with the ideas of statistical thinking before you involve yourself with the details of statistical methods. It is easy to get caught in the trap of doing rather than understanding, of asking how rather than why. You can’t do statistics unless you learn the methods, but you must not get so caught up in the details of the methods that you lose sight of what they mean. Doing and thinking, method and meaning, will compete for your attention throughout this course.

**In this chapter, you will learn the basic ideas of**

- exploring data—uncovering and summarizing patterns
- making inferences from data—deciding whether an observed feature of the data could reasonably be attributed to chance alone

These ideas will remain key components of the statistical concepts you’ll develop and study throughout this course.
Robert Martin was one of 50 people working in the engineering department of Westvaco's envelope division. One spring, Westvaco's management went through five rounds of planning for a reduction in their workforce. In Round 1, they eliminated 11 positions, and they eliminated 9 more in Round 2. By the time the layoffs ended, after all five rounds, only 22 of the 50 workers had kept their jobs. The average age in the department had fallen from 48 to 46.

After Martin, age 54, was laid off, he sued Westvaco for age discrimination. Display 1.1 shows the data provided by Westvaco to Martin’s lawyers. The statistical analysis in the lawsuit used all 50 employees in the engineering department of the envelope division, with separate analyses for salaried and hourly workers. Each row in Display 1.1 corresponds to one worker, and each column corresponds to a characteristic of the worker: job title, whether hourly or salaried, month and year of birth, month and year of hire, and age at birthday in 1991. The next-to-last column (Round) tells how the worker fared in the downsizing: 1 means chosen for layoff in Round 1 of planning for the reduction in force, 2 means chosen in Round 2, and so on for Rounds 3, 4, and 5; 0 means “not chosen for layoff.”

The subjects (or objects) of statistical examination often are called cases. In the rows in Display 1.1, the cases are individual Westvaco employees. Their characteristics, in the columns, are the variables. If you pick a row and read across, you find information about a single case. (For example, Robert Martin, in Row 44, was salaried, was born in September 1937, was hired in October 1967, was chosen for layoff in Round 2, and turned 54 in 1991.) Although reading across might seem the natural way to read the table, in statistics you will often find it useful to pick a column and read down. This gives you information about a single variable as you range through all the cases. For example, pick Age, read down the column, and notice the variability in the ages. It is variability like this—the fact that individuals differ—that can make it a challenge to see patterns in data and figure out what they mean.

Imagine: If there had been no variability—if all the workers had been of just two ages, say, 30 and 50, and Westvaco had laid off all the 50-year-olds and kept all the 30-year-olds—the conclusion would be obvious and there would be no need for statistics. But real life is more subtle than that. The ages of the laid-off workers varied, as did the ages of the workers retained. Statistical methods were designed to cope with such variability. In fact, you might define statistics as the science of learning from data in the presence of variability.

Although the bare fact that the ages vary is easy to see in the data table, the pattern of those ages is not so easy to see. This pattern—what the values are and how often each occurs—is their distribution. In order to see that pattern, a graph is better than a table. The dot plot in Display 1.2 shows the distribution of the ages of the 36 salaried employees who worked in the engineering department just before the layoffs began.

[To learn how to create a dot plot on your calculator, see Calculator Note 1A.]
<table>
<thead>
<tr>
<th>Row</th>
<th>Job Title</th>
<th>Pay</th>
<th>Birth</th>
<th>Hire</th>
<th>Age at Birthday in 1991</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Engineering Clerk</td>
<td>H</td>
<td>9 66</td>
<td>7 89</td>
<td>25</td>
</tr>
<tr>
<td>2</td>
<td>Engineering Tech II</td>
<td>H</td>
<td>4 53</td>
<td>8 78</td>
<td>38</td>
</tr>
<tr>
<td>3</td>
<td>Engineering Tech II</td>
<td>H</td>
<td>10 35</td>
<td>7 65</td>
<td>56</td>
</tr>
<tr>
<td>4</td>
<td>Secretary to Engin Manag</td>
<td>H</td>
<td>2 43</td>
<td>9 66</td>
<td>48</td>
</tr>
<tr>
<td>5</td>
<td>Engineering Tech II</td>
<td>H</td>
<td>8 38</td>
<td>9 74</td>
<td>53</td>
</tr>
<tr>
<td>6</td>
<td>Engineering Tech II</td>
<td>H</td>
<td>8 36</td>
<td>3 60</td>
<td>55</td>
</tr>
<tr>
<td>7</td>
<td>Engineering Tech II</td>
<td>H</td>
<td>1 32</td>
<td>2 63</td>
<td>59</td>
</tr>
<tr>
<td>8</td>
<td>Parts Crib Attendant</td>
<td>H</td>
<td>11 69</td>
<td>10 89</td>
<td>22</td>
</tr>
<tr>
<td>9</td>
<td>Engineering Tech II</td>
<td>H</td>
<td>5 36</td>
<td>4 77</td>
<td>55</td>
</tr>
<tr>
<td>10</td>
<td>Engineering Tech II</td>
<td>H</td>
<td>8 27</td>
<td>12 51</td>
<td>64</td>
</tr>
<tr>
<td>11</td>
<td>Technical Secretary</td>
<td>H</td>
<td>5 36</td>
<td>11 73</td>
<td>55</td>
</tr>
<tr>
<td>12</td>
<td>Engineering Tech II</td>
<td>H</td>
<td>2 36</td>
<td>4 62</td>
<td>55</td>
</tr>
<tr>
<td>13</td>
<td>Engineering Tech II</td>
<td>H</td>
<td>9 58</td>
<td>11 76</td>
<td>33</td>
</tr>
<tr>
<td>14</td>
<td>Engineering Tech II</td>
<td>H</td>
<td>7 56</td>
<td>5 77</td>
<td>35</td>
</tr>
<tr>
<td>15</td>
<td>Customer Serv Engineer</td>
<td>S</td>
<td>4 30</td>
<td>9 66</td>
<td>61</td>
</tr>
<tr>
<td>16</td>
<td>Customer Serv Engr Assoc</td>
<td>S</td>
<td>2 62</td>
<td>5 88</td>
<td>29</td>
</tr>
<tr>
<td>17</td>
<td>Design Engineer</td>
<td>S</td>
<td>12 43</td>
<td>9 67</td>
<td>48</td>
</tr>
<tr>
<td>18</td>
<td>Design Engineer</td>
<td>S</td>
<td>3 37</td>
<td>6 74</td>
<td>54</td>
</tr>
<tr>
<td>19</td>
<td>Design Engineer</td>
<td>S</td>
<td>3 36</td>
<td>2 78</td>
<td>55</td>
</tr>
<tr>
<td>20</td>
<td>Design Engineer</td>
<td>S</td>
<td>1 31</td>
<td>3 67</td>
<td>60</td>
</tr>
<tr>
<td>21</td>
<td>Engineering Assistant</td>
<td>S</td>
<td>6 60</td>
<td>7 86</td>
<td>31</td>
</tr>
<tr>
<td>22</td>
<td>Engineering Associate</td>
<td>S</td>
<td>2 57</td>
<td>4 85</td>
<td>34</td>
</tr>
<tr>
<td>23</td>
<td>Engineering Manager</td>
<td>S</td>
<td>2 32</td>
<td>11 63</td>
<td>59</td>
</tr>
<tr>
<td>24</td>
<td>Machine Designer</td>
<td>S</td>
<td>9 59</td>
<td>3 90</td>
<td>32</td>
</tr>
<tr>
<td>25</td>
<td>Packaging Engineer</td>
<td>S</td>
<td>3 38</td>
<td>11 83</td>
<td>53</td>
</tr>
<tr>
<td>26</td>
<td>Prod Spec—Printing</td>
<td>S</td>
<td>12 44</td>
<td>11 74</td>
<td>47</td>
</tr>
<tr>
<td>27</td>
<td>Proj Eng—Elec</td>
<td>S</td>
<td>9 43</td>
<td>4 71</td>
<td>48</td>
</tr>
<tr>
<td>28</td>
<td>Project Engineer</td>
<td>S</td>
<td>7 49</td>
<td>9 73</td>
<td>42</td>
</tr>
<tr>
<td>29</td>
<td>Project Engineer</td>
<td>S</td>
<td>8 43</td>
<td>4 64</td>
<td>48</td>
</tr>
<tr>
<td>30</td>
<td>Project Engineer</td>
<td>S</td>
<td>6 34</td>
<td>8 81</td>
<td>57</td>
</tr>
<tr>
<td>31</td>
<td>Supv Engineering Serv</td>
<td>S</td>
<td>4 54</td>
<td>6 72</td>
<td>37</td>
</tr>
<tr>
<td>32</td>
<td>Supv Machine Shop</td>
<td>S</td>
<td>11 37</td>
<td>3 64</td>
<td>54</td>
</tr>
<tr>
<td>33</td>
<td>Chemist</td>
<td>S</td>
<td>8 22</td>
<td>4 54</td>
<td>69</td>
</tr>
<tr>
<td>34</td>
<td>Design Engineer</td>
<td>S</td>
<td>9 38</td>
<td>12 87</td>
<td>53</td>
</tr>
<tr>
<td>35</td>
<td>Engineering Associate</td>
<td>S</td>
<td>2 61</td>
<td>9 85</td>
<td>30</td>
</tr>
<tr>
<td>36</td>
<td>Machine Designer</td>
<td>S</td>
<td>2 39</td>
<td>4 85</td>
<td>52</td>
</tr>
<tr>
<td>37</td>
<td>Machine Parts Cont—Supv</td>
<td>S</td>
<td>10 28</td>
<td>8 53</td>
<td>63</td>
</tr>
<tr>
<td>38</td>
<td>Prod Specialist</td>
<td>S</td>
<td>9 27</td>
<td>10 43</td>
<td>64</td>
</tr>
<tr>
<td>39</td>
<td>Project Engineer</td>
<td>S</td>
<td>7 25</td>
<td>9 59</td>
<td>66</td>
</tr>
<tr>
<td>40</td>
<td>Chemist</td>
<td>S</td>
<td>12 30</td>
<td>10 52</td>
<td>61</td>
</tr>
<tr>
<td>41</td>
<td>Design Engineer</td>
<td>S</td>
<td>4 60</td>
<td>5 89</td>
<td>31</td>
</tr>
<tr>
<td>42</td>
<td>Electrical Engineer</td>
<td>S</td>
<td>11 49</td>
<td>3 86</td>
<td>42</td>
</tr>
<tr>
<td>43</td>
<td>Machine Designer</td>
<td>S</td>
<td>3 35</td>
<td>12 68</td>
<td>56</td>
</tr>
<tr>
<td>44</td>
<td>Machine Parts Cont Coor</td>
<td>S</td>
<td>9 37</td>
<td>10 67</td>
<td>54</td>
</tr>
<tr>
<td>45</td>
<td>VH Prod Specialist</td>
<td>S</td>
<td>5 35</td>
<td>9 55</td>
<td>56</td>
</tr>
<tr>
<td>46</td>
<td>Printing Coordinator</td>
<td>S</td>
<td>2 41</td>
<td>1 62</td>
<td>50</td>
</tr>
<tr>
<td>47</td>
<td>Prod Dev Engineer</td>
<td>S</td>
<td>6 59</td>
<td>11 85</td>
<td>32</td>
</tr>
<tr>
<td>48</td>
<td>Prod Specialist</td>
<td>S</td>
<td>7 32</td>
<td>1 55</td>
<td>59</td>
</tr>
<tr>
<td>49</td>
<td>VH Prod Specialist</td>
<td>S</td>
<td>3 42</td>
<td>4 62</td>
<td>49</td>
</tr>
<tr>
<td>50</td>
<td>Engineering Associate</td>
<td>S</td>
<td>8 68</td>
<td>5 89</td>
<td>23</td>
</tr>
</tbody>
</table>

Display 1.2 Ages of the salaried workers. (Each dot represents a worker; the age is shown by the position of the dot along the scale below it.)

Display 1.2 provides some useful information about the variability in the ages, but by itself doesn't tell anything about possible age discrimination in the layoffs. For that, you need to distinguish between those salaried workers who lost their jobs and those who didn't. The dot plot in Display 1.3, which shows those laid off and those retained, provides weak evidence for Martin's case. Those laid off generally were older than those who kept their jobs, but the pattern isn't striking.

Display 1.3 Salaried workers: ages of those laid off and those retained.

Display 1.3 shows that most salaried workers who were laid off were age 50 or older. However, this alone doesn't support Martin's case because most of the workers were age 50 or older to begin with.

One way to proceed is to make a summary table. The table shown here classifies the salaried workers according to age and whether they were laid off or retained. (Using 50 as the dividing age between “younger” and “older” is a somewhat arbitrary, but reasonable, decision.)

<table>
<thead>
<tr>
<th></th>
<th>Laid Off</th>
<th>Retained</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under 50</td>
<td>6</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>50 or Older</td>
<td>12</td>
<td>8</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>18</td>
<td>18</td>
<td>36</td>
</tr>
</tbody>
</table>

To decide whether Martin has a case, compare the proportion of salaried workers under 50 who were laid off (6 out of 16, or 0.375) with the proportion of those 50 or older who were laid off (12 out of 20, or 0.60). These proportions are quite different, an argument in favor of Martin.

Looking at the layoffs of the salaried workers round by round provides further evidence in favor of Martin. The dot plots in Display 1.4 show the ages of the salaried workers laid off and retained by round. These new dot plots use different
symbols for laid-off workers and for retained workers. For example, in the top dot plot, the open circles represent the salaried workers whose jobs did not survive Round 1. In this round, the four oldest workers were laid off, but only one worker under age 50 was laid off. In the second round, one of the two oldest employees was laid off. But then this pattern stopped. Again, the evidence favors Martin but is far from conclusive.

Display 1.4  Salaried workers: ages of those laid off (open circles) and those retained (solid dots) in each round.

You might feel as if the analysis so far ignores important facts, such as worker qualifications. That’s true. However, the first step is to decide whether, based on the data in Display 1.1, older workers were more likely to be laid off. If not, Martin’s case fails. If so, it is then up to Westvaco to justify its actions.

DISCUSSION

Exploring the Martin v. Westvaco Data

D1. Suppose you were the judge in the Martin v. Westvaco case. How would you use the information in Display 1.1 to decide whether Westvaco tended to lay off older workers in disproportionate numbers (for whatever reason)?

D2. Display 1.5 (on the next page) is like Display 1.3 except that it gives data for the hourly workers. Compare the plots for the hourly and salaried workers. Which provides stronger evidence in support of Martin’s claim of age discrimination?
D3. Whenever you think you have a message from data, you should be careful not to jump to conclusions. The patterns in the Westvaco data might be “real”—they reflect age discrimination on the part of management. On the other hand, the patterns might be the result of chance—management wasn’t discriminating on the basis of age but simply by chance happened to lay off a larger percentage of older workers. What’s your opinion about the Westvaco data: Do the patterns seem “real”—too strong to be explained by chance?

D4. The analysis up to this point ignores important facts such as worker qualifications. Suppose Martin makes a convincing case that older workers were more likely to be laid off. It is then up to Westvaco to justify its actions. List several specific reasons Westvaco might give to justify laying off a disproportionate number of older workers.

**Summary 1.1: Data Exploration**

Data exploration, or exploratory analysis, is a purposeful investigation to find patterns in data, using tools such as tables and graphs to display those patterns and statistical concepts such as distributions and averages to summarize them.

- A table display, with cases listed in rows and variables in columns, helps you look at how variables differ from case to case.

- The distribution of a variable tells you the set of values that the variable takes on, together with how often each value occurs.

- A dot plot, which shows the values of a variable along a number line, provides you with a visual display of the distribution of the variable and gives you a sense of how large or small the values are, which values occur most often, how spread out the values are, and whether any values appear to be unusually large or small.

Statistics involves coping with variability, so you have to understand the causes of that variability before you can draw informed conclusions. All the features of data exploration that you have investigated here will be important when you move on to the inference phase of a statistical investigation.
Practice

Practice problems help you master basic concepts and computations. Throughout this textbook, you should work all the practice problems for each topic you want to learn. The answers to all practice problems are given in the back of the book.

Exploring the Martin v. Westvaco Data

P1. Construct a dot plot similar to Display 1.3, comparing the ages of hourly workers who lost their jobs during Rounds 1–3 to the ages of hourly workers who still had their jobs at the end of Round 3. How do the ages differ?

P2. This summary table classifies the hourly workers according to age and whether they were laid off or retained.

<table>
<thead>
<tr>
<th>Age</th>
<th>Laid Off</th>
<th>Retained</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under 50</td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>50 or Older</td>
<td>7</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td>4</td>
<td>14</td>
</tr>
</tbody>
</table>

a. What proportion of hourly workers under age 50 were laid off? Were not laid off?
b. What proportion of laid-off hourly workers were under age 50? Were age 50 or older?
c. What two proportions should you compute and compare in order to decide whether older hourly workers were disproportionately laid off? Make these computations and give your conclusion.
d. Compare this table with the table for the salaried workers on page 6. For which group does the evidence more strongly favor Martin’s case?

P3. Display 1.6 shows layoffs and retentions by round for hourly workers. (There is no plot for Round 5 because no hourly workers were chosen for layoff in that round.) Compare the pattern for the hourly workers with the pattern for the salaried workers in Display 1.4 on page 7. For which group does the evidence more strongly favor Martin’s case?

<table>
<thead>
<tr>
<th>Age</th>
<th>Round 1</th>
<th>Round 2</th>
<th>Round 3</th>
<th>Round 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>. . .</td>
<td>. . .</td>
<td>. . .</td>
<td>. . .</td>
</tr>
<tr>
<td>30</td>
<td>. .</td>
<td>. .</td>
<td>. .</td>
<td>. .</td>
</tr>
<tr>
<td>40</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Display 1.6 Hourly workers: ages of those laid off (open circles) and those retained (solid dots) in each round.

Exercises

E1. This summary table classifies salaried workers as to whether they were laid off and their age, this time using 40 as the cutoff between younger and older workers.

<table>
<thead>
<tr>
<th>Age</th>
<th>Laid Off</th>
<th>Retained</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under 40</td>
<td>4</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>40 or Older</td>
<td>14</td>
<td>13</td>
<td>27</td>
</tr>
<tr>
<td>Total</td>
<td>18</td>
<td>18</td>
<td>36</td>
</tr>
</tbody>
</table>

a. What proportion of workers age 40 or older were laid off? What proportion of laid-off workers were age 40 or older?
b. What proportion of workers under age 40 were laid off? What proportion were not laid off?
c. What two proportions should you compute and compare in order to
decide whether older workers were disproportionately laid off? Make these computations and give your conclusion.

d. Compare this table with the table for the salaried workers on page 6, where 50 was the age cutoff. If you were Martin’s lawyer, would you present a table using 40 or 50 as the cutoff?

E2. Explore whether hourly workers at Westvaco were more likely than salaried workers to lose their jobs.

a. Start by constructing a summary table to display the relevant data.

b. Compute two proportions that will allow you to make this comparison.

c. What do you conclude from comparing the proportions?

E3. Twenty-two workers kept their jobs. Explore whether the age distributions are similar for the hourly and salaried workers who kept their jobs.

a. Show the two age distributions on a pair of dot plots that have the same scale. How do these distributions differ?

b. Do your dot plots in part a support a claim that Westvaco was more inclined to keep older workers if they were salaried rather than hourly?

E4. Consider these three facts from your work in this section:

• Salaried workers were more likely to keep their jobs than were hourly workers.

• Older workers were more likely to be laid off than were younger workers.

• Older workers were more likely to be salaried than were younger workers.

Putting these three facts together, what can you conclude? Is this evidence in favor of Martin’s case, or does it help Westvaco?

E5. Refer to Display 1.1 on page 5.

a. Create a summary table whose five cases are Round 1 through Round 5 and whose three variables are total number of employees laid off in that round, number of employees laid off in that round who were 40 or older, and percentage laid off in that round who were 40 or older.

b. Describe any patterns you find in the table and what you think they might mean.

E6. “Last hired, first fired” is shorthand for “When you have to downsize, start by laying off the newest person, then the person hired next before that, and work back in reverse order of seniority.” (The person who’s been working longest will be the last to be laid off.) Examine the Westvaco data.

a. How was seniority related to the decisions about layoffs in the engineering department at Westvaco?

b. What explanation(s) can you suggest for any patterns you find?

E7. Many tables in the media are arranged with cases as rows and variables as columns. For Displays 1.7 and 1.8 in parts a and b, identify the cases and the variables. Then compute the values missing from each table.
Suppose you are studying the effects of poverty and plan to construct a data set whose cases are the villages in Bolivia. Name some variables that you might study.

1.2 Discrimination in the Workplace: Inference

Overall, the exploratory work on the Westvaco data set in Section 1.1 shows that older workers were more likely than younger workers to be laid off and were laid off earlier. One of the main arguments in the court case, along the lines set out in D3, was about what those patterns mean: Can you infer from the patterns that Westvaco has some explaining to do, or are they the sort of patterns that tend to happen even in the absence of discrimination?

A comprehensive analysis of *Martin v. Westvaco* will have to wait for its reappearance among the case studies of Chapter 12, when you'll be more familiar with the concepts and tools of statistics. For now, you can get a pretty good idea of how the analysis goes by working with a subset of the data.
The ages of the ten hourly workers involved in Round 2 of the layoffs, arranged from youngest to oldest, were 25, 33, 35, 38, 48, 55, 55, 55, 56, and 64. The three workers who were laid off were ages 55, 55, and 64. Display 1.9 shows these data on a dot plot.

To simplify the statistical analysis to come, it helps to “condense” the data into a single number, called a **summary statistic**. One possible summary statistic is the **average**, or **mean**, age of the three workers who lost their jobs:

\[
\frac{55 + 55 + 64}{3} = 58 \text{ years}
\]

Knowing what to make of the data requires balancing two points of view. On one hand, the pattern in the data is pretty striking. Of the five workers under age 50, all kept their jobs. Of the five who were 55 or older, only two kept their jobs. On the other hand, the number of workers involved is small: only three out of ten. Should you take seriously a pattern involving so few cases? Imagine two people taking sides in an argument that was at the center of the statistical part of the Martin case.

**Martin**: Look at the pattern in the data. All three of the workers laid off were much older than the average age of all workers. That’s evidence of age discrimination.

**Westvaco**: Not so fast! You’re looking at only ten workers total, and only three positions were eliminated. Just one small change and the picture would be entirely different. For example, suppose it had been the 25-year-old instead of the 64-year-old who was laid off. Switch the 25 and the 64, and you get a totally different set of averages. (Ages in red are those selected for layoff.)
See! Make just one small change, and the average age of the three who were laid off is lower than the average age of the others.

<table>
<thead>
<tr>
<th>Laid Off</th>
<th>Retained</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Data</td>
<td>58.0</td>
</tr>
<tr>
<td>Altered Data</td>
<td>45.0</td>
</tr>
</tbody>
</table>

**Martin:** Not so fast yourself! Of all the possible changes, you picked the one most favorable to your side. If you’d switched one of the 55-year-olds who got laid off with the 55-year-old who kept his or her job, the averages wouldn’t change at all. Why not compare what actually happened with all the possibilities?

**Westvaco:** What do you mean?

**Martin:** Start with the ten workers, and pick three at random. Do this over and over, to see what typically happens, and compare the actual data with the results. Then we’ll find out how likely it is that the average age of those laid off would be 58 or greater.

The dialogue between Martin and Westvaco describes one age-neutral method for choosing which workers to lay off: Pick three workers completely at random, with all sets of three having the same chance to be chosen.

**DISCUSSION**

**Picking Workers at Random**

D5. If you pick three of the ten ages at random, do you think you are likely to get an average age of 58 or greater?

D6. If the probability of getting an average age of 58 or greater turns out to be small, does this favor Martin or Westvaco?

Activity 1.2a shows you how to estimate the probability of getting an average age of 58 years or greater if you choose three workers at random. You will use simulation, a procedure in which you set up a model of a chance process (drawing three ages out of a box) that copies, or simulates, a real situation (selecting three employees at random to lay off).

**ACTIVITY 1.2a**

**By Chance or by Design?**

**What you’ll need:** paper or 3 × 5 cards, a box or other container

Let’s test the process suggested by Martin’s advocate.

1. Create a model of a chance process. Write each of the ten ages on identical pieces of paper or 3 × 5 cards, and put the ten cards in a box. Mix them thoroughly, draw out three (the ones to be laid off), and record the ages.

(continued)
Chapter 1 Statistical Reasoning: Investigating a Claim of Discrimination

(continued)

2. **Compute a summary statistic.** Compute the average of the three numbers in your sample to one decimal place.

3. **Repeat the process.** Repeat steps 1 and 2 nine more times.

4. **Display the distribution.** Pool your results with the rest of your class and display the average ages on a dot plot.

5. **Estimate the probability.** Count the number of times your class computed an average age of 58 years or greater. Estimate the probability that simply by chance the average age of those chosen would be 58 years or greater.

6. **Interpret your results.** What do you conclude from your class’s estimate in step 5?

[See Calculator Note 1B to learn how to do this kind of simulation with your calculator.]

Your simulation was completely age-neutral. All sets of three workers had exactly the same chance of being selected for layoff, regardless of age. The simulation tells you what results are reasonable to expect from that sort of age-blind process.

Shown here are the first 4 of 200 repetitions from such a simulation. (The ages in red are those selected for layoff.) The average ages of the workers selected for layoff—42.7, 48.0, 42.7, and 37.0—are highlighted by the red dots in the distribution of all 200 repetitions in Display 1.10.

*Display 1.10* Results of 200 repetitions: the distribution of the average age of the three workers chosen for layoff by chance alone.
Out of 200 repetitions, only 10, or 5%, gave an average age of 58 or greater. So it is not at all likely that simply by chance you'd pick workers as old as the three Westvaco picked. Did the company discriminate? There's no way to tell from the numbers alone—Westvaco might have a good explanation. On the other hand, if your simulation had told you that an average of 58 or greater is easy to get by chance alone, then the data would provide no evidence of discrimination and Westvaco wouldn't need to explain.

To better understand how this logic applies to *Martin v. Westvaco*, imagine a realistic argument between the advocates for each side.

**Martin:** Look at the pattern in the data. All three of the workers laid off were much older than average.

**Westvaco:** So what? You could get a result like that just by chance. If chance alone can account for the pattern, there's no reason to ask us for any other explanation.

**Martin:** Of course you *could* get this result by chance. The question is whether it's easy or hard to do so. If it's easy to get an average as large as 58 by drawing at random, I'll agree that we can't rule out chance as one possible explanation. But if an average that large is really hard to get from random draws, we agree that it's not reasonable to say that chance alone accounts for the pattern. Right?

**Westvaco:** Right.

**Martin:** Here are the results of my simulation. If you look at the three hourly workers laid off in Round 2, the probability of getting an average age of 58 or greater by chance alone is only 5%. And if you do the same computations for the entire engineering department, the probability is a lot lower, about 1%. What do you say to that?

**Westvaco:** Well... I'll agree that it's really hard to get an average age that extreme simply by chance, but that by itself still doesn't prove discrimination.

**Martin:** No, but I think it leaves you with some explaining to do!

In the actual case, Martin and Westvaco reached a settlement out of court before the case went to trial.

The logic you've just seen is basic to all statistical inference, but it's not easy to understand. In fact, it took mathematicians centuries to come up with the ideas. It wasn't until the 1920s that a brilliant British biological scientist and mathematician, R. A. Fisher, realized that results of agricultural experiments may be analyzed in a way similar to that in Activity 1.2a to see whether observed differences should be attributed to chance alone or to treatment. Calculus, in contrast, was first understood in 1665. Precisely because it is so important, the logic of using randomization as a basis for statistical inference will be seen over and over again throughout this book. You'll have lots of time to practice with it.
**Key Steps in a Simulation**

Here is a summary of the steps in a simulation:

1. **Model.** Set up a model in which chance is the only cause of being selected.
   
   In Activity 1.2a the model for an age-neutral chance process was to draw three numbers at random from the set of ten ages. You put the ages of the workers who could be laid off in a box and selected three by random draw.

2. **Repetition.** Repeat the process.
   
   In the activity, you repeated the process of drawing ages many times.

3. **Distribution.** Display the distribution of the summary statistics, and determine how likely the actual result or one even more extreme would be.
   
   In the activity, you used the average age to summarize the results, although other summary statistics also could be used. For each repetition, you computed the average age of those laid off and plotted that average on your dot plot. The simulation showed that the chance of an average age of 58 or greater was only about 0.05.

4. **Conclusion.** If the probability is small (and the definition of small will vary depending on the situation), conclude that some explanation other than just chance should be considered. If the probability isn't small, conclude that you can reasonably attribute the result to chance alone.

   In Round 2 of the *Martin v. Westvaco* case, the small probability of 0.05 (1 chance out of 20) meant that Westvaco had some explaining to do. However, the Round 2 evidence alone wouldn't have been enough to serve as evidence of discrimination in a court case (which requires a probability of 0.025 or less).

**DISCUSSION**

**The Logic of Inference**

D7. Why must you estimate the probability of getting an average age of 58 or greater rather than the probability of getting an average age of 58?

D8. *How unlikely is “too unlikely”?* The probability you estimated in Activity 1.2a is in fact exactly equal to 0.05. In a typical court case, a probability of 0.025 or less is required to serve as evidence of discrimination.

   a. Did the Round 2 layoffs of hourly workers in the *Martin* case meet the court requirement?

   b. If the probability in the *Martin* case had been 0.01 instead of 0.05, how would that have changed your conclusions? 0.10 instead of 0.05?


   a. Explain why the evidence—19 heads in 20 flips—makes it hard to believe the coin is fair.

   b. Design and carry out a simulation to estimate how unusual this result would be if the coin were fair.
Summary 1.2: Inference

Inference is a statistical procedure that involves deciding whether an event can reasonably be attributed to chance or whether you should look for—and perhaps investigate—some other explanation. In the Martin case, you used inference to determine whether the relatively high average age of the laid-off hourly employees in Round 2 could reasonably be due to chance.

Simulation is a useful device for inference.

- First, you set up a model of a process in which chance is the only factor influencing the outcomes.
- The next stage is repetition—you repeat the process.
- Then you plot the distribution of the summary statistics in order to determine how likely the actual result or one even more extreme would be.
- Finally, you reach—or infer—a conclusion. If the probability of getting a summary statistic as extreme as that from your actual data is small, conclude that chance isn't a reasonable explanation. If the probability isn't small, conclude that you can reasonably attribute the result to chance alone.

In the Martin case, the probability was about 0.05, which was considered small enough to warrant asking for an explanation from Westvaco but not small enough to present in court as clear evidence of discrimination.

Practice

The Logic of Inference

P4. Suppose three workers were laid off from a set of ten whose ages were the same as those of the hourly workers in Round 2 in the Martin case. This time, however, the ages of those laid off were 48, 55, and 55.

25 33 35 38 48 55 55 56 64

a. Use the dot plot in Display 1.10 on page 14 to estimate the probability of getting an average age as large as or larger than that of those laid off in this situation.

b. What would your conclusion be if Westvaco had laid off workers of these three ages?

P5. At the beginning of Round 1, there were 14 hourly workers. Their ages were 22, 25, 33, 35, 38, 48, 53, 55, 55, 55, 55, 56, 59, and 64. After the layoffs were complete, the ages of those left were 25, 38, 48, and 56. Think about how you would repeat Activity 1.2a using these data.

a. What is the average age of the ten workers laid off?

b. Describe a simulation for finding the distribution of the average age of ten workers laid off at random.

c. The results of 200 repetitions from a simulation are shown in Display 1.11. Suppose 10 workers are picked at random for layoff from the 14 hourly workers. Make a rough estimate of the probability of getting, just by chance, the same or larger average age as that of the workers who actually were laid off (from part a).

d. Does this analysis provide evidence in Martin's favor?
Exercises

E9. Revisit the idea of the simulation in Activity 1.2a, this time for all 14 hourly workers and using a different summary statistic. Use as your summary statistic the number of hourly workers laid off who were 40 or older. The ages listed here are those of the hourly workers, with the ages of those laid off in red. Note that, of the ten hourly workers laid off by Westvaco, seven were age 40 or older.

22 25 33 35 38 48 53 55 55 55 55 56 59 64

a. Write the 14 ages on 14 slips of paper and draw 10 at random to be chosen for layoff. How many of the 10 are age 40 or older?
b. The dot plot in Display 1.12 shows the results of 50 repetitions of this simulation. Estimate the probability that, by chance, seven or more of the ten hourly workers who were laid off would be age 40 or older.
c. Do you conclude that the proportion of laid-off workers age 40 or older could reasonably be due to chance alone, or should Westvaco be asked for an explanation?

E10. The ages of the ten hourly workers left after Round 1 are given here. The ages of the four workers laid off in Rounds 2 and 3 are shown in red. Their average age is 57.25.

25 33 35 38 48 55 55 55 56 64

a. Describe how to simulate the chance of getting an average age of 57.25 or more using the methods of Activity 1.2a.
b. Perform your simulation once and compute the average age of the four hourly workers laid off.
c. The dot plot in Display 1.13 shows the results of 200 repetitions of this simulation. What is your estimate of the probability of getting an average age as great as or greater than Westvaco did if four workers are picked at random for layoff in Rounds 2 and 3 from the ten hourly workers remaining after Round 1?
d. What is your conclusion?

E11. Mrs. Garcia was not happy when she found that her baker had raised the price of a loaf of bread—and she let him know it. However, she did buy her usual three loaves of bread. They seemed a little light, so she asked that they be weighed and that the other eight loaves the baker could have given her also be weighed. The other eight loaves weighed 14, 15, 15, 16, 16, 17, 17, and 18 ounces. The three loaves Mrs. Garcia was given weighed...
14, 15, and 16 ounces. The baker claimed that he picked the three loaves at random. In Display 1.14, each dot represents the average weight of a random sample of three loaves.

Which of these conclusions should Mrs. Garcia draw?

A. Because the probability of getting an average weight as low as or lower than that of Mrs. Garcia’s three loaves is small, Mrs. Garcia should not be suspicious that the baker deliberately gave her lighter loaves.

B. Because the probability of getting an average weight as low as or lower than that of Mrs. Garcia’s three loaves is fairly large, Mrs. Garcia should be suspicious that the baker deliberately gave her lighter loaves.

E12. Snow in July? You have spent some time in Oz. You think the date back in Kansas is July 4, but you can’t be sure because days might not have the same length in Oz as on Earth. A friendly tornado puts you and your dog Toto down in Kansas. However, you see snow falling (data). Which of these inferences should you make?

A. If this is Kansas, it is very unlikely to be snowing on July 4. Therefore, this probably isn’t Kansas.

B. If it is July 4, it is very unlikely to be snowing in Kansas. Therefore, this probably isn’t July 4.

C. If it is snowing in Kansas on July 4, it is time to go back to Oz.

D. If this is Kansas and it is July 4, it probably isn’t really snowing.

E13. For some situations, instead of using simulation, it is possible to find exact probabilities by counting equally likely outcomes. Suppose only two out of the ten hourly workers had been laid off in Round 2 and that those two workers were ages 55 and 64, with an average age of 59.5. It is straightforward, though tedious, to list all possible pairs of workers who might have been chosen. Here’s the beginning of a systematic listing. The first nine outcomes include the 25-year-old and one other. The next eight outcomes include the 33-year-old and one other, but not the 25-year-old because the pair {25, 33} was already counted.
a. How many possible pairs are there? (Don't list them all!)

b. How many pairs give an average age of 59.5 or greater? (Do list them.)

c. If the pair is chosen completely at random, then all possibilities are equally likely and the probability of getting an average age of 59.5 or greater equals the number of pairs with an average of 59.5 or more divided by the total number of possible pairs. What is the probability?

d. Is the evidence of age discrimination strong or weak?

E14. In this exercise, you will follow the same steps as in E13 to find the probability of getting an average age of 58 or greater when drawing three hourly workers at random in Round 2. The number of ways to pick three different workers from ten to lay off is

\[ \binom{10}{3} = \frac{10!}{3!(10-3)!} = 120 \]

[See Calculator Note 1C to learn how to compute numbers of combinations.]

a. List the ways that give an average age of 58 or greater.

b. Compute the probability of getting an average age of 58 or greater when three workers are selected for layoff at random.

c. How does this probability compare to the results of your class simulation in Activity 1.2a? Why do the two probabilities differ (if they do)?

Chapter Summary

In this chapter, you explored the data from an actual case of alleged age discrimination, looking for evidence you considered relevant. You then saw how to use statistical reasoning to test the strength of the evidence: Are the patterns in the data solid enough to support Martin's claim of age discrimination, or are they the sort that you would expect to occur even if there was no discrimination? Along the way you made a substantial start at learning many of the most important statistical terms and concepts: distribution, cases and variables, summary statistic, simulation, and how to determine whether the result from the real-life situation can reasonably be attributed to chance alone or whether an explanation is called for.

You have practiced both thinking like a statistician and reporting your results like a statistician. Throughout this textbook, you will be asked to justify your answers in the real-world context. This includes stating assumptions, giving appropriate plots and computations, and writing a conclusion in context.

The last chapter of this book includes a final look at the Martin case.

Review Exercises

E15. A teacher had two statistics classes, and students could enroll in either the earlier class or the later class. The final grades in the courses are given here.

<table>
<thead>
<tr>
<th>Earlier class:</th>
<th>99</th>
<th>95</th>
<th>69</th>
<th>91</th>
<th>79</th>
<th>67</th>
<th>64</th>
<th>54</th>
<th>68</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>47</td>
<td>53</td>
<td>86</td>
<td>100</td>
<td>95</td>
<td>45</td>
<td>41</td>
<td>59</td>
<td>66</td>
</tr>
<tr>
<td>Later class:</td>
<td>84</td>
<td>68</td>
<td>94</td>
<td>77</td>
<td>88</td>
<td>75</td>
<td>88</td>
<td>91</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>61</td>
<td>97</td>
<td>75</td>
<td>37</td>
<td>82</td>
<td>62</td>
<td>49</td>
<td>43</td>
<td>93</td>
</tr>
</tbody>
</table>

E20. Chapter 1 Statistical Reasoning: Investigating a Claim of Discrimination
a. Display these data on dot plots so that you can compare the two classes.

b. What conclusion can you draw from the dot plots? Could the difference between the two classes reasonably be attributed to chance, or do you think the teacher should look for an explanation?

E16. Refer to the data in E15.

a. Make a table that divides the course grades into “fail” (less than 60) and “pass” (60 or more) for the two classes.

b. What proportion of the students in the earlier class passed? What proportion of students who passed were in the earlier class? What proportion of students passed overall?

c. What two proportions should you compute and compare in order to decide whether a disproportionate number of “passing” students enrolled in the earlier class? Make these computations and give your conclusion.

E17. This table classifies the Westvaco workers by whether they were laid off and whether they were under age 50 or were age 50 or older.

<table>
<thead>
<tr>
<th></th>
<th>Laid Off</th>
<th>Retained</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under 50</td>
<td>9</td>
<td>13</td>
<td>22</td>
</tr>
<tr>
<td>50 or Older</td>
<td>19</td>
<td>9</td>
<td>28</td>
</tr>
<tr>
<td>Total</td>
<td>28</td>
<td>22</td>
<td>50</td>
</tr>
</tbody>
</table>

Choose the best conclusion to draw from this table.

A. The table supports a claim of age discrimination because most of the people who were laid off were 50 or older.

B. The table supports a claim of age discrimination because a larger percentage of the people age 50 or older were laid off than people under 50.

C. The table supports a claim of age discrimination because a larger percentage of the laid-off people were age 50 or older than were under 50.

D. The table does not support a claim of age discrimination because the number of people under age 50 who were laid off is equal to the number of people age 50 or older who were retained.

E. The table does not support a claim of age discrimination because a larger percentage of people were 50 or older to begin with.

E18. Display 1.15 contains information about the planets in our solar system. The radius of each planet is given in miles, and the temperature is the average at the surface. What are the cases? What are the variables?

<table>
<thead>
<tr>
<th>Planet</th>
<th>Radius</th>
<th>Moons</th>
<th>Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mercury</td>
<td>1,516</td>
<td>0</td>
<td>332°F</td>
</tr>
<tr>
<td>Venus</td>
<td>3,760</td>
<td>0</td>
<td>67°F</td>
</tr>
<tr>
<td>Earth</td>
<td>3,963</td>
<td>1</td>
<td>59°F</td>
</tr>
<tr>
<td>Mars</td>
<td>2,111</td>
<td>2</td>
<td>−82°F</td>
</tr>
<tr>
<td>Jupiter</td>
<td>44,423</td>
<td>63</td>
<td>−163°F</td>
</tr>
<tr>
<td>Saturn</td>
<td>37,449</td>
<td>56</td>
<td>−218°F</td>
</tr>
<tr>
<td>Uranus</td>
<td>15,882</td>
<td>27</td>
<td>−323°F</td>
</tr>
<tr>
<td>Neptune</td>
<td>15,388</td>
<td>13</td>
<td>−330°F</td>
</tr>
</tbody>
</table>

Display 1.15 Data about planets in our solar system. [Source: solarsystem.nasa.gov.]

E19. Earlier you studied the summary table of salaried workers classified according to age and to whether they were laid off or retained, using 50 as the dividing age. That table is shown again here.

<table>
<thead>
<tr>
<th></th>
<th>Laid Off</th>
<th>Retained</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under 50</td>
<td>6</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>50 or Older</td>
<td>12</td>
<td>8</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>18</td>
<td>18</td>
<td>36</td>
</tr>
</tbody>
</table>

The proportion of those under age 50 who were laid off (6 out of 16, or 0.375) is smaller than the proportion of those age 50 or older who were laid off (12 out of 20, or 0.60). The key question, however, is “Is the actual 12 versus 6 split of those laid off consistent with selecting workers at random for layoff?”

a. Which of these demonstrations would help Martin’s case?

A. Showing that it’s not unusual to get 12 or more older workers if 18 workers are selected at random for layoff
B. Showing that it’s pretty unusual to get 12 or more older workers if 18 workers are selected at random for layoff

b. To investigate this situation using simulation, follow these steps once and record your results.

1. Make 36 identical white cards. Label 20 cards with an O for “50 or over” and 16 cards with a U for “under 50.”
2. Mix the cards in a bag and select the 18 to be laid off at random.
3. Count the number of O’s among the 18 selected.

c. Display 1.16 shows the results of a computer simulation of 200 repetitions conducted according to the rules given in part b. Where would your result from part b be placed on this dot plot?

d. From this simulation, a total of 26 out of 200 repetitions gave counts of “50 or older” that were 12 or more. What percentage of the repetitions is this? Is this percentage small enough to cast serious doubt on a claim that those laid off were chosen by chance?

E20. The Eastbanko Company had fifteen workers before laying off five of them. The ages of the fifteen workers were 22, 23, 25, 31, 34, 36, 37, 40, 41, 43, 44, 50, 55, 55, and 60, with the ages of the five laid-off workers in bold.

The dot plot in Display 1.17 gives the results of a simulation with 600 repetitions for the average age of five of these workers chosen for layoff at random. Each dot represents the average of five ages, rounded down to a whole number.
a. Estimate the probability that if workers were selected by chance alone for layoff, the average age of those laid off would be as large as or larger than the average age of those in the actual layoffs.

b. If the 60-year-old sues for age discrimination, would Eastbanko have some explaining to do?

E21. In E9, you conducted a simulation to estimate the probability that, just by chance, seven or more of the ten hourly workers who were laid off would be age 40 or older. The ages of the fourteen hourly workers were 22, 25, 33, 35, 38, 48, 53, 55, 55, 55, 55, 56, 59, and 64.

a. How many ways can 10 workers be selected from 14 workers for layoff?

b. If there are a total of 10 layoffs, what numbers of older workers would it have been possible to lay off?

c. Using your calculator, find the number of ways that you can lay off
   i. seven older workers and three younger workers
   ii. eight older workers and two younger workers
   iii. nine older workers and one younger worker

d. What is the probability that you will get 7 or more workers age 40 or older if you select 10 of the 14 workers completely at random for layoff?

E22. Refer to your reasoning in E14, where you computed the probability that the three workers laid off in Round 2 would have an average age of 58 or greater. Describe how your reasoning and conclusions would be different if the workers’ ages were 25, 33, 35, 38, 48, 55, 55, 55, and 55, and the three workers chosen for layoff were all age 55. Is the evidence stronger or weaker for Martin in this situation than in E14?

E23. The Society for the Preservation of Wild Gnus held a raffle last week and sold 50 tickets. The two lucky participants whose tickets were drawn received all-day passes to the Wild Gnu Park in Florida. But there was a near riot when the winners were announced—both winning tickets belonged to society president Filbert Newman’s cousins. After some intense questioning by angry ticket holders, it was determined that only 4 of the 50 tickets belonged to Newman’s cousins and the other 46 tickets belonged to people who were not part of his family. Newman’s final comment to the press was “Hey kids, I guess we were just lucky. Deal with it.”

One member of the Gnu Society was taking a statistics class and decided to deal with it by simulating the drawing. He put 50 tickets in a bowl; 4 of the tickets were marked “C” for “cousin” and 46 were marked “N” for “not a cousin.” The statistics student drew two tickets at random and kept track of the number of cousins picked. After doing this 1000 times, the student found that 844 draws resulted in two N’s, 149 in one N and one C, and only 7 in two C’s.

a. Use the results of the simulation to estimate the probability that, in a fair drawing, both winning tickets would be held by Newman’s cousins.

b. Using the probability you estimated in part a, write a short paragraph that the statistics student can send to other members of the Gnu Society.

c. Is it possible that Newman’s cousins won the prizes by chance alone? Explain.

d. Using reasoning like that in E13 and E14, compute the exact probability that, in a fair drawing, both winning tickets would be held by Newman’s cousins.
AP1. This plot shows the ages of the part-time and full-time students who receive financial aid at a small college. Which of the following is a conclusion about students at this college that cannot be drawn from the plot alone?

- Part-time students who receive financial aid tend to be older than full-time students who receive financial aid.
- A larger proportion of part-time students than full-time students receive financial aid.
- The oldest student receiving financial aid is a full-time student.
- No student under age 18 receives financial aid.
- More part-time students than full-time students receive financial aid.

AP2. This table classifies hourly and salaried workers as to whether they were laid off. Do the data support a claim that hourly workers are being treated unfairly?

<table>
<thead>
<tr>
<th>Laid Off</th>
<th>Not Laid Off</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hourly</td>
<td>31</td>
<td>31</td>
</tr>
<tr>
<td>Salaried</td>
<td>24</td>
<td>14</td>
</tr>
<tr>
<td>Total</td>
<td>55</td>
<td>45</td>
</tr>
</tbody>
</table>

- Yes, because most people laid off were hourly workers.
- Yes, because a bigger proportion of hourly workers were laid off than salaried workers.
- No, because half of hourly workers were laid off and half were not.
- No, because more than half of workers were hourly and less than half salaried.

AP3. This table shows the number of male and female applicants who applied and were either admitted to or rejected from a graduate program. What proportion of admitted applicants were female?

<table>
<thead>
<tr>
<th></th>
<th>Admitted</th>
<th>Rejected</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>17</td>
<td>33</td>
<td>50</td>
</tr>
<tr>
<td>Female</td>
<td>8</td>
<td>12</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>45</td>
<td>70</td>
</tr>
</tbody>
</table>

- Yes, because half of hourly workers were laid off, but more than half of salaried workers were laid off.

AP4. For the data in AP3, in order to determine if there is evidence to continue investigating whether the graduate admissions process discriminates against females, a study takes a random sample of 25 out of the 70 applicants to be the “admitted” group. The proportion of females in the sample was computed. This process was repeated for a total of 50 random samples and the results are graphed below. What is the best conclusion to draw from this simulation?

- The actual proportion of females among those admitted is very near the center of this distribution, so there is no evidence of discrimination.
- The actual proportion of females among those admitted is very near the center of this distribution, so there is strong evidence of discrimination in favor of female applicants.
The actual proportion of females among those admitted is very near the center of this distribution, so there is strong evidence of discrimination against female applicants.

The actual proportion of females among those admitted is quite a bit above the center of this distribution, so there is strong evidence of discrimination in favor of female applicants.

The actual proportion of females among those admitted is quite a bit below the center of this distribution, so there is strong evidence of discrimination against female applicants.

Investigative Tasks

AP5. People with asthma often use an inhaler to help open up their lungs and breathing passages. A pharmaceutical company has come up with a new compound to put in the inhaler that, they believe, will open up the lungs of the user even more than the standard compound tends to do. Ten volunteers with asthma are randomly split into two groups: one group uses the new compound B and the other uses the standard compound A. The measurements listed in Display 1.18 are the increase in lung capacity (in liters) 1 hour after the use of the inhaler.

<table>
<thead>
<tr>
<th>Compound A</th>
<th>Compound B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.03</td>
<td>1.11</td>
</tr>
<tr>
<td>0.45</td>
<td>1.01</td>
</tr>
<tr>
<td>0.32</td>
<td>0.44</td>
</tr>
<tr>
<td>0.64</td>
<td>1.41</td>
</tr>
<tr>
<td>1.29</td>
<td>1.04</td>
</tr>
</tbody>
</table>

Display 1.18 Increase in lung capacity, in liters, 1 hour after use of an inhaler containing a compound.

a. From simply studying the data in the table, do you think compound B does better than compound A in increasing lung capacity?

b. Construct dot plots for compounds A and B. Does it now appear that compound B tends to give larger measurements than compound A?

c. Find the average increase in lung capacity for compound A and for compound B. When you compare these means, does it look to you as if compound B is better than compound A at opening up the lungs?

AP6. Refer to AP5. Your task now is to see whether the observed difference in the means of each treatment group reasonably could be attributed to chance alone.

a. Place the ten measurements on separate slips of paper and mix them in a bag. Select five at random to play the role of the A treatment group; the other five play the role of the B treatment group. This time you will use as your summary statistic the difference between the means of each treatment group. Calculate this difference, mean (compound B) — mean (compound A), for your sample.

b. The dot plot in Display 1.19 shows the results of 50 repetitions of this simulation. Compute the difference between the means for the actual data. Locate this difference on the dot plot. How many simulated differences exceed the actual difference? What proportion?

Display 1.19 Results of 50 repetitions: the distribution of the difference between the means of two randomly selected groups of five values.

c. In light of this simulation, do you think it is reasonable to attribute the actual difference to chance alone? Explain.
What does the distribution of female heights look like? Statistics gives you the tools to visualize and describe large sets of data.
“Raw” data—a long list of values—are hard to make sense of. Suppose, for example, that you are applying to the University of Michigan at Ann Arbor and wonder how your SAT I math score of 650 compares with those of the students who attend that university. If all you have are raw data—a list of the SAT I math scores of the 25,000 students at the University of Michigan—it would take a lot of time and effort to make sense of the numbers.

Suppose instead that you read the summary in the university’s guide, which says that “the middle 50% of the SAT math scores were between 630 and 720, with half above 680 and half below.” Now you know that your score of 650, though in the bottom half of the scores, is not far from the center value of 680 and is above the bottom quarter.

Notice that the summary of the scores gives you two different kinds of information for the middle 50%: the **center**, 680, and the **spread**, from 630 to 720. Often that’s all you need, especially if the **shape** of the distribution is one of a few standard shapes you’ll learn about in this chapter.

These three features—shape, center, and spread—can sometimes take you a surprisingly long way in data analysis. For example, in Chapter 1 you did a simulation to answer this question: If you choose three people at random from a set of ten people and compute the average age of the ones you choose, how likely is it that you get an average of 58 years or more? Generally, you don’t need to do all this work! Using shape, center, and spread, you can get an answer without doing a simulation. This remarkable fact, which first began to come to light in the late 1600s, helped make statistical inference possible in the 20th century before the age of computers. In the next several chapters, you’ll learn how to make good use of this fact.

**In this chapter, you begin your systematic study of distributions by learning how to**

- make and interpret different kinds of plots
- describe the shapes of distributions
- choose and compute a measure of center
- choose and compute a useful measure of spread (variability)
- work with the normal distribution
Visualizing Distributions: Shape, Center, and Spread

Summaries simplify. In fact, summaries sometimes can oversimplify, which means it is important to know when to use summaries and which summaries to use. Often the right choice depends on the shape of your distribution. To help you build your visual intuition about how shape and summaries are related, this first section of the chapter introduces various shapes and asks you to estimate some summary values visually. Later sections will tell you how to compute summary values numerically.

Distributions come in a variety of shapes. Four of the most common shapes are illustrated in the rest of this section.

Uniform (Rectangular) Distributions

The plot to the left shows the shape of a uniform or rectangular distribution, in which all values occur equally often. How uniform is a sample of values taken from a uniform distribution? In the next activity, you will find out.

ACTIVITY 2.1a
Distributing Digits

What you’ll need: one page from a phone book for each member of the class, and a box of slips of paper, with one slip for each member of your class, half labeled “phone book” and half “fake it”

1. Suppose your class made a dot plot of the last digits of every phone number in the phone book. (This would take a very long time!) Sketch what you think this plot would look like.

2. Draw a slip of paper from the box.
   - If the slip you drew says “phone book,” use the page from the phone book, start at a random spot, and write down the last digit of each of the next 30 phone numbers. Using a full sheet of paper, plot your 30 digits on a dot plot, using a scale like the one here. Use big dots so that they can be seen from across the room.

   - If the slip you drew says “fake it,” don’t use the page from the phone book but instead make up and plot 30 digits on a dot plot using a scale

(continued)
like the one on the previous page. Try to make the distribution look like
the digits might have come from the phone book.

Write your name on the front of the plot, but not which method you
used. Don’t consult with other students while you are doing this step.

3. Post the dot plots around the room and compare. Which plots are you
confident came from the phone book? Which are you confident came from
made-up digits? (Don’t say anything about your own plot.)

4. Find your plot and write a large P (for “phone book”) or F (for “faked it”)
on the front. Check your predictions from step 3. What differences do you
see in the two groups of plots?

5. In this activity, it was important that you sampled from the phone book
in such a way that all digits were equally likely to occur. Why did step 1
specify that you use the last digit of the phone number and not, say,
the first?

The number of births per month in a year is another set of data you might expect
to be fairly uniform. Or, is there a reason to believe that more babies are born in
one month than in another? Display 2.1 shows a table and plot of U.S. births (in
thousands) for 2003.

<table>
<thead>
<tr>
<th>Month</th>
<th>Births (in thousands)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>330</td>
</tr>
<tr>
<td>2</td>
<td>307</td>
</tr>
<tr>
<td>3</td>
<td>337</td>
</tr>
<tr>
<td>4</td>
<td>330</td>
</tr>
<tr>
<td>5</td>
<td>347</td>
</tr>
<tr>
<td>6</td>
<td>337</td>
</tr>
<tr>
<td>7</td>
<td>364</td>
</tr>
<tr>
<td>8</td>
<td>360</td>
</tr>
<tr>
<td>9</td>
<td>360</td>
</tr>
<tr>
<td>10</td>
<td>354</td>
</tr>
<tr>
<td>11</td>
<td>320</td>
</tr>
<tr>
<td>12</td>
<td>344</td>
</tr>
</tbody>
</table>

Display 2.1 An example of a (roughly) uniform distribution:

[Source: Centers for Disease Control and Prevention.]

The plot shows that there is actually little change from month to month;
that is, we see a roughly uniform distribution of births across the months. To
summarize this distribution, you might write “The distribution of births is
roughly uniform over the months January through December, with about
340,000 births per month.”
Computers and many calculators generate random numbers between 0 and 1 with a uniform distribution. Display 2.2 shows a dot plot of 1000 random numbers generated by statistical software. There is some variability in the frequencies, but, as expected, about 20% of the random numbers fall between, for example, 0.2 and 0.4. [See Calculator Note 2A to learn how to create a distribution of random numbers using your calculator.]

Display 2.2  Dot plot of 1000 random numbers from a uniform distribution. Each dot represents two points.

### Uniform Distributions

D1. Think of other situations that you would expect to be uniform distributions
   a. over the days of the week
   b. over the digits 0, 1, 2, \ldots, 9

D2. Think of situations that you would expect to be very nonuniform distributions
   a. over the months of the year
   b. over the days of the month
   c. over the digits 0, 1, 2, \ldots, 9
   d. over the days of the week

### Normal Distributions

Activity 2.1b introduces one of the most important common shapes of distributions and one of the common ways this shape is produced. What happens when different people measure the same distance or the same feature of very similar objects? In the activity, you’ll measure a tennis ball with a ruler, but the results you get will reflect what happens even if you use very precise instruments under carefully controlled conditions. For example, a 10-gram platinum weight is used for calibration of scales all across the United States. When scientists at the National Institute of Standards and Technology use an analytical balance for the weight’s weekly weighing, they face a similar challenge due to variability.

### Activity 2.1b  Measuring Diameters

**What you’ll need:** a tennis ball, a ruler with a centimeter scale

1. With your partner, plan a method for measuring the diameter of the tennis ball with the centimeter ruler.

(continued)
2. Using your method, make two measurements of the diameter of your tennis ball to the nearest millimeter.

3. Combine your data with those of the rest of the class and make a dot plot. Speculate first, about the shape you expect for the distribution.

4. *Shape.* What is the approximate shape of the plot? Are there clusters and gaps or unusual values (outliers) in the data?

5. *Center and spread.* Choose two numbers that seem reasonable for completing this sentence: “Our typical diameter measurement is about —?—, give or take about —?—.” (More than one reasonable set of choices is possible.)

6. Discuss some possible reasons for the variability in the measurements. How could the variability be reduced? Can the variability be eliminated entirely?

The measurements of the diameter of a tennis ball taken by your class in Activity 2.1b probably were not uniform. More likely, they piled up around some central value, with a few measurements far away on the low side and a few far away on the high side. This common bell shape has an idealized version—the normal distribution—that is especially important in statistics.

Pennies minted in the United States are supposed to weigh 3.110 g, but a tolerance of 0.130 g is allowed in either direction. Display 2.3 shows a plot of the weights of 100 pennies.

The smooth curve superimposed on the graph of the pennies is an example of a normal curve. No real-world example matches the curve perfectly, but many plots of data are approximately normal. The idealized normal shape is perfectly symmetric—the right side is a mirror image of the left side, as in Display 2.4. There is a single peak, or mode, at the line of symmetry, and the curve drops off smoothly on both sides, flattening toward the x-axis but never quite reaching it and stretching infinitely far in both directions. On either side of the mode are inflection points, where the curve changes from concave down to concave up.
Display 2.4  A normal curve, showing the line of symmetry, mode, mean, inflection points, and standard deviation (SD).

You should use the **mean** (or average) to describe the center of a normal distribution. The mean is the value at the point where the line of symmetry intersects the $x$-axis. You should use the **standard deviation**, $SD$ for short, to describe the spread. The $SD$ is the horizontal distance from the mean to an inflection point.

It is difficult to locate inflection points, especially when curves are drawn by hand. A more reliable way to estimate the standard deviation is to use areas. For a normal curve, 68% (roughly) of the total area under the curve is between the vertical lines through the two inflection points. In other words, the interval between one standard deviation on either side of the mean accounts for roughly 68% of the area under the normal curve.

**Example: Averages of Random Samples**

Display 2.5 shows the distribution of average ages computed from 200 sets of five workers chosen at random from the ten hourly workers in Round 2 of the Westvaco case discussed in Chapter 1. Notice that, apart from the bumpiness, the shape is roughly normal. Estimate the mean and standard deviation.

**Display 2.5**  Distribution of average age for groups of five workers drawn at random.
Solution

The curve in the display has center at 47, and the middle 68% of dots fall roughly between 43 and 51. Thus, the estimated mean is 47, and the estimated standard deviation is 4. A typical random sample of five workers has average age 47 years, give or take about 4 years.

[You can graph a normal curve on your calculator by specifying the mean and standard deviation. See Calculator Note 2B.]

In this section, you’ve seen the three most common ways normal distributions arise in practice:

- through variation in measurements (diameters of tennis balls)
- through natural variation in populations (weights of pennies)
- through variation in averages computed from random samples (average ages)

All three scenarios are common, which makes the normal distribution especially important in statistics.

DISCUSSION

Normal Distributions

D3. Determine these summary statistics visually.

a. Estimate the mean and standard deviation of the penny weight data in Display 2.3, and use your estimates to write a summary sentence.

b. Estimate the mean and standard deviation of your class data from Activity 2.1b.

Skewed Distributions

Both the uniform (rectangular) and normal distributions are symmetric. That is, if you smooth out minor bumps, the right side of the plot is a mirror image of the left side. Not all distributions are symmetric, however. Many common distributions show bunching at one end and a long tail stretching out in the other direction. These distributions are called skewed. The direction of the tail tells whether the distribution is skewed right (tail stretches right, toward the high values) or skewed left (tail stretches left, toward the low values).

The dot plot in Display 2.6 shows the weights, in pounds, of 143 wild bears. It is skewed right (toward the higher values) because the tail of the distribution stretches out in that direction. In everyday conversation, you might describe the two parts of
the distribution as “normal” and “abnormal.” Usually, bears weigh between about 50 and 250 lb (this part of the distribution even looks approximately normal), but if someone shouts “Abnormal bear loose!” you should run for cover—that unusual bear is likely to be big! The “unusualness” of the distribution is all in one direction.

Often the bunching in a skewed distribution happens because values “bump up against a wall”—either a minimum that values can’t go below, such as 0 for measurements and counts, or a maximum that values can’t go above, such as 100 for percentages. For example, the distribution in Display 2.7 shows the grade-point averages of college students (mostly first-year students and sophomores) taking an introductory statistics course at the University of Florida. It is skewed left (toward the smaller values). The maximum grade-point average is 4.0, for all A’s, so the distribution is bunched at the high end because of this wall. A GPA of 0.0 wouldn’t be called a wall, even though GPAs can’t go below 0.0, because the values aren’t bunched up against it. The skew is to the left: An unusual GPA would be one that is low compared to most GPAs of students in the class.

Use the median along with the lower and upper quartiles to describe the center and spread of a skewed distribution.

Because there is no line of symmetry in a skewed distribution, the ideas of center and spread are not as clear-cut as they are for a normal distribution. To get around this problem, typically you should use the median to describe the center of a skewed distribution. To estimate the median from a dot plot, locate the value that divides the dots into two halves, with equal numbers of dots on either side.

You should use the lower and upper quartiles to indicate spread. The lower quartile is the value that divides the lower half of the distribution into two halves, with equal numbers of dots on either side. The upper quartile is the value that divides the upper half of the distribution into two halves, with equal numbers of dots on either side. The three values—lower quartile, median, and upper quartile—divide the distribution into quarters. This allows you to describe a distribution as in the introduction to this chapter: “The middle 50% of the SAT math scores were between 630 and 720, with half above 680 and half below.”

**Example: Median and Quartiles for Bear Weights**

Divide the bears’ weights in Display 2.8 into four groups of equal size, and estimate the median and quartiles. Write a short summary of this distribution.

**Solution**

There are 143 dots in Display 2.8, so there are about 71 or 72 dots in each half and 35 or 36 in each quarter. The value that divides the dots in half is about 155 lb. The values that divide the two halves in half are roughly 115 and 250. Thus, the middle 50% of the bear weights are between about 115 and 250, with half above about 155 and half below.
2.1 Visualizing Distributions: Shape, Center, and Spread

![Weight distribution graph](image)

**Display 2.8** Estimating center and spread for the weights of bears.

---

**DISCUSSION**

**Skewed Distributions**

D4. Decide whether each distribution described will be skewed. Is there a wall that leads to bunching near it and a long tail stretching out away from it? If so, describe the wall.

a. the sizes of islands in the Caribbean
b. the average per capita incomes for the nations of the United Nations
c. the lengths of pant legs cut and sewn to be 32 in. long
d. the times for 300 university students of introductory psychology to complete a 1-hour timed exam
e. the lengths of reigns of Japanese emperors

D5. Which would you expect to be the more common direction of skew, right or left? Why?

**Bimodal Distributions**

Many distributions, including the normal distribution and many skewed distributions, have only one peak (*unimodal*), but some have two peaks (*bimodal*) or even more. When your distribution has two or more obvious peaks, or modes, it is worth asking whether your cases represent two or more groups. For example, Display 2.9 shows the life expectancies of females from countries on two continents, Europe and Africa.

![Bimodal distribution graph](image)

**Display 2.9** Life expectancy of females by country on two continents. [Source: Population Reference Bureau, *World Population Data Sheet*, 2005.]
Europe and Africa differ greatly in their socioeconomic conditions, and the life expectancies reflect those conditions. If you make a separate plot for each of the two continents, the two peaks become essentially one peak in each plot, as in Display 2.10.

Although it makes sense to talk about the center of the distribution of life expectancies for Europe or for Africa, notice that it doesn’t really make sense to talk about “the” center of the distribution for both continents together. You could possibly tell the locations of the two peaks, but finding the reason for the two modes and separating the cases into two distributions communicates even more.

**Other Features: Outliers, Gaps, and Clusters**

An unusual value, or **outlier**, is a value that stands apart from the bulk of the data. Outliers always deserve special attention. Sometimes they are mistakes (a typing mistake, a measuring mistake), sometimes they are atypical for other reasons (a really big bear, a faulty lab procedure), and sometimes unusual features of the distribution are the key to an important discovery.

In the late 1800s, John William Strutt, third Baron Rayleigh (English, 1842–1919), was studying the density of nitrogen using samples from the air outside his laboratory (from which known impurities were removed) and samples produced by a chemical procedure in his lab. He saw a pattern in the results that you can observe in the plot of his data in Display 2.11.

Lord Rayleigh saw two clusters separated by a gap. (There is no formal definition of a **gap** or a **cluster**; you have to use your best judgment about them. For example, some people call a single outlier a cluster of one; others don’t. You
also could argue that the value at the extreme right is an outlier, perhaps because of a faulty measurement.

When Rayleigh checked the clusters, it turned out that the ten values to the left had all come from the chemically produced samples and the nine to the right had all come from the atmospheric samples. What did this great scientist conclude? The air samples on the right might be denser because of something in them besides nitrogen. This hypothesis led him to discover inert gases in the atmosphere.

**Summary 2.1: Visualizing Distributions**

Distributions have different shapes, and different shapes call for different summaries.

- If your distribution is uniform (rectangular), it's often enough simply to tell the range of the set of values and the approximate frequency with which each value occurs.
- If your distribution is normal (bell-shaped), you can give a good summary with the mean and the standard deviation. The mean lies at the center of the distribution, and the standard deviation ($SD$) is the horizontal distance from the center to the points of inflection, where the curvature changes. To estimate the $SD$, find the distance on either side of the mean that defines the interval enclosing about 68% of the cases.
- If your distribution is skewed, you can give the values (median and quartiles) that divide the distribution into fourths.
- If your distribution is bimodal, reporting a single center isn't useful. One reasonable summary is to locate the two peaks. However, it is even more useful if you can find another variable that divides your set of cases into two groups centered at the two peaks.

Later in this chapter, you will study the various measures of center and spread in more detail and learn how to compute them.

**Practice**

Practice problems help you master basic concepts and computations. Throughout this textbook, you should work all the practice problems for each topic you want to learn. The answers to all practice problems are given in the back of the book.

**Uniform (Rectangular) Distributions**

P1. This diagram shows a uniform distribution on $[0, 2]$, the interval from 0 through 2.

```
0 2
```

- a. What value divides the distribution in half, with half the numbers below that value and half above?
- b. What values divide the distribution into quarters?
- c. What values enclose the middle 50% of the distribution?
- d. What percentage of the values lie between 0.4 and 0.7?
- e. What values enclose the middle 95% of the distribution?
P2. The plot in Display 2.12 gives the number of deaths in the United States per month in 2003. Does the number of deaths appear to be uniformly distributed over the months? Give a verbal summary of the way deaths are distributed over the months of the year.


Normal Distributions

P3. For each of the normal distributions in Display 2.13, estimate the mean and standard deviation visually, and use your estimates to write a verbal summary of the form “A typical SAT score is roughly (mean), give or take (SD) or so.”

Display 2.13  Four distributions that are approximately normal.

Skewed Distributions

P4. Estimate the median and quartiles for the distribution of GPAs in Display 2.7 on page 34. Then write a verbal summary of the same form as in the example.

P5. Match each plot in Display 2.14 with its median and quartiles (the set of values that divide the area under the curve into fourths).

a. SAT math scores
b. ACT scores
c. heights of women attending college
d. single-season batting averages for professional baseball players in the 1910s

d. 35, 50, 65
e. 25, 50, 75

Display 2.14  Five distributions with different shapes.
Exercises

E1. Describe each distribution as bimodal, skewed right, skewed left, approximately normal, or roughly uniform.

a. ages of all people who died last year in the United States
b. ages of all people who got their first driver's license in your state last year
c. SAT scores for all students in your state taking the test this year
d. selling prices of all cars sold by General Motors this year

E2. Describe each distribution as bimodal, skewed right, skewed left, approximately normal, or roughly uniform.

a. the incomes of the world’s 100 richest people
b. the birthrates of Africa and Europe
c. the heights of soccer players on the last Women's World Cup championship team
d. the last two digits of telephone numbers in the town where you live
e. the length of time students used to complete a chapter test, out of a 50-minute class period

E4. The U.S. Environmental Protection Agency’s National Priorities List Fact Book tells the number of hazardous waste sites for each U.S. state and territory. For 2006, the numbers ranged from 1 to 138, the middle 50% of the values were between 11 and 32, half the values were above 18, and half were below 18. Sketch what the distribution might look like. [Source: U.S. Environmental Protection Agency, www.epa.gov, 2006.]

E5. The dot plot in Display 2.15 gives the ages of the officers who attained the rank of colonel in the Royal Netherlands Air Force.

a. What are the cases? Describe the variables.
b. Describe this distribution in terms of shape, center, and spread.
c. What kind of wall might there be that causes the shape of the distribution? Generate as many possibilities as you can.

![The 2003 Women's World Cup Championship team, from Germany](image)

Display 2.15 Ages of colonels. Each dot represents two points. [Source: Data and Story Library at Carnegie-Mellon University, lib.stat.cmu.edu.]
E6. The dot plot in Display 2.16 shows the distribution of the number of inches of rainfall in Los Angeles for the seasons 1899–1900 through 1999–2000.

Display 2.16  Los Angeles rainfall. [Source: National Weather Service.]

a. What are the cases? Describe the variables.
b. Describe this distribution in terms of shape, center, and spread.
c. What kind of wall might there be that causes the shape of the distribution? Generate as many possibilities as you can.

E7. The distribution in Display 2.17 shows measurements of the strength in pounds of 22s yarn (22s refers to a standard unit for measuring yarn strength). What is the basic shape of this distribution? What feature makes it uncharacteristic of distributions with that shape?

Display 2.17  Strength of yarn. [Source: Data and Story Library at Carnegie-Mellon University, lib.stat.cmu.edu.]

E8. Sketch a normal distribution with mean 0 and standard deviation 1. You will study this standard normal distribution in Section 2.5.

E9. Make up a scenario (name the cases and variables) whose distribution you would expect to be

a. skewed right because of a wall. What is responsible for the wall?
b. skewed left because of a wall. What is responsible for the wall?

E10. The plot in Display 2.18 shows the last digit of the Social Security numbers of the students in a statistics class. Describe this distribution.

Display 2.18  Last digit of a sample of Social Security numbers.

E11. Although a uniform distribution gives a reasonable approximation of the actual distribution of births over months (Display 2.1 on page 29), you can “blow up” the graph to see departures from the uniform pattern, as in Display 2.19. Do these deviations from the uniform shape form their own pattern, or do they appear haphazard? If you think there's a pattern, describe it.

Display 2.19  A “blow up” of the distribution of births over months, showing departures from the uniform pattern.
E12. Draw a graph similar to that in Display 2.19 for the data on deaths in the United States listed in Display 2.20, and summarize what you find.

<table>
<thead>
<tr>
<th>Month</th>
<th>Deaths (in thousands)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>224</td>
</tr>
<tr>
<td>2</td>
<td>200</td>
</tr>
<tr>
<td>3</td>
<td>212</td>
</tr>
<tr>
<td>4</td>
<td>194</td>
</tr>
<tr>
<td>5</td>
<td>197</td>
</tr>
<tr>
<td>6</td>
<td>193</td>
</tr>
<tr>
<td>7</td>
<td>192</td>
</tr>
<tr>
<td>8</td>
<td>188</td>
</tr>
<tr>
<td>9</td>
<td>192</td>
</tr>
<tr>
<td>10</td>
<td>204</td>
</tr>
<tr>
<td>11</td>
<td>197</td>
</tr>
<tr>
<td>12</td>
<td>230</td>
</tr>
</tbody>
</table>

Display 2.20 Deaths in the United States, 2003. [Source: Centers for Disease Control and Prevention.]

E13. How do countries compare with respect to the value of the goods they produce? Display 2.21 shows gross domestic product (GDP) per capita, a measure of the total value of all goods and services produced divided by the number of people in a country, and the average number of people per room in housing units, a measure of crowdedness, for a selection of countries in Asia, Europe, and North America. You’ll analyze these data in parts a–d.

<table>
<thead>
<tr>
<th>Country</th>
<th>Per Capita GDP (in U.S. dollars)</th>
<th>Average Number of People per Room</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austria</td>
<td>31,187</td>
<td>0.7</td>
</tr>
<tr>
<td>Azerbaijan</td>
<td>853</td>
<td>2.1</td>
</tr>
<tr>
<td>Belgium</td>
<td>29,257</td>
<td>0.6</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>2,533</td>
<td>1.0</td>
</tr>
<tr>
<td>Canada</td>
<td>27,097</td>
<td>0.5</td>
</tr>
<tr>
<td>China</td>
<td>1,100</td>
<td>1.1</td>
</tr>
<tr>
<td>Croatia</td>
<td>6,398</td>
<td>1.2</td>
</tr>
<tr>
<td>Cyprus</td>
<td>16,038</td>
<td>0.6</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>8,834</td>
<td>1.0</td>
</tr>
<tr>
<td>Finland</td>
<td>31,069</td>
<td>0.8</td>
</tr>
<tr>
<td>France</td>
<td>29,222</td>
<td>0.7</td>
</tr>
<tr>
<td>Germany</td>
<td>29,137</td>
<td>0.5</td>
</tr>
<tr>
<td>Hungary</td>
<td>8,384</td>
<td>0.8</td>
</tr>
<tr>
<td>India</td>
<td>555</td>
<td>2.7</td>
</tr>
<tr>
<td>Iraq</td>
<td>594</td>
<td>1.5</td>
</tr>
<tr>
<td>Israel</td>
<td>18,101</td>
<td>1.2</td>
</tr>
<tr>
<td>Japan</td>
<td>33,819</td>
<td>0.8</td>
</tr>
<tr>
<td>Korea, Republic of</td>
<td>11,059</td>
<td>1.1</td>
</tr>
<tr>
<td>Kuwait</td>
<td>13,641</td>
<td>1.7</td>
</tr>
<tr>
<td>Netherlands</td>
<td>31,759</td>
<td>0.7</td>
</tr>
<tr>
<td>Norway</td>
<td>48,881</td>
<td>0.6</td>
</tr>
<tr>
<td>Pakistan</td>
<td>498</td>
<td>3.0</td>
</tr>
<tr>
<td>Poland</td>
<td>5,355</td>
<td>1.0</td>
</tr>
<tr>
<td>Portugal</td>
<td>14,645</td>
<td>0.7</td>
</tr>
<tr>
<td>Romania</td>
<td>2,550</td>
<td>1.3</td>
</tr>
<tr>
<td>Serbia-Montenegro</td>
<td>1,843</td>
<td>1.2</td>
</tr>
<tr>
<td>Slovakia</td>
<td>6,019</td>
<td>1.2</td>
</tr>
<tr>
<td>Sri Lanka</td>
<td>913</td>
<td>2.2</td>
</tr>
<tr>
<td>Sweden</td>
<td>33,925</td>
<td>0.5</td>
</tr>
<tr>
<td>Switzerland</td>
<td>43,486</td>
<td>0.6</td>
</tr>
<tr>
<td>Syria</td>
<td>1,497</td>
<td>2.0</td>
</tr>
<tr>
<td>Turkey</td>
<td>3,418</td>
<td>1.3</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>30,355</td>
<td>0.5</td>
</tr>
<tr>
<td>United States</td>
<td>36,924</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Display 2.21 Per capita GDP and crowdedness for a selection of countries. [Source: United Nations, unstats.un.org.]
A dot plot of the per capita GDP data is shown in Display 2.22.

Display 2.22  Dot plot of per capita GDP.

a. Describe this distribution in terms of shape, center, and spread.
b. Which two countries have the highest per capita GDP? Do they appear to be outliers?
c. A rather large gap appears near the middle of the distribution. Which of the two clusters formed by this gap contains mostly Western European and North American countries? In what part of the world are most of the countries in the other cluster?

d. Is it surprising to find clusters and gaps in data that measure an aspect of the economies of the countries?

E14. The dot plot in Display 2.23 gives a look at how the countries listed in Display 2.21 compare in terms of the crowdedness of their residents.

Display 2.23  Dot plot of crowdedness.

a. Describe this distribution in terms of shape, center, and spread.
b. Which countries appear to be outliers? Are they the same as the countries that appeared to be outliers for the per capita GDP data?
c. Where on the dot plot is the cluster that contains mostly Western European and North American countries?

2.2 Graphical Displays of Distributions

As you saw in the previous section, the best way to summarize a distribution often depends on its shape. To see the shape, you need a suitable graph. In this section, you’ll learn how to make and interpret three kinds of plots for quantitative variables (dot plot, histogram, and stemplot) and one plot for categorical variables (bar chart).

Cases and Variables, Quantitative and Categorical

Pet cats typically live about 12 years, but some have been known to live 28 years. Is that typical of domesticated predators? What about domesticated nonpredators, such as cows and guinea pigs? What about wild mammals? The rhinoceros, a nonpredator, lives an average of 15 years, with a maximum of about 45 years. The grizzly bear, a wild predator, lives an average of 25 years, with a maximum of about 50 years. Do meat-eaters typically outlive vegetarians in the wild? Often you can find answers to questions like these in a plot of the data.

Many of the examples in this section are based on the data about mammals in Display 2.24. Each row (type of mammal) is a case. As you learned in Chapter 1, the cases in a data set are the people, cities, mammals, or other items being studied.
Measurements and other properties of the cases are organized into columns, with one column for each variable. Thus, *average longevity* and *speed* are variables, and, for example, 30 mi/h is the value of the variable *speed* for the case grizzly bear.

<table>
<thead>
<tr>
<th>Mammal</th>
<th>Gestation Period (days)</th>
<th>Average Longevity (years)</th>
<th>Maximum Longevity (years)</th>
<th>Speed (mi/h)</th>
<th>Wild (1 = yes; 0 = no)</th>
<th>Predator (1 = yes; 0 = no)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baboon</td>
<td>187</td>
<td>20</td>
<td>45</td>
<td>*</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Bear, grizzly</td>
<td>225</td>
<td>25</td>
<td>50</td>
<td>30</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Beaver</td>
<td>105</td>
<td>5</td>
<td>50</td>
<td>*</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Bison</td>
<td>285</td>
<td>15</td>
<td>40</td>
<td>*</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Camel</td>
<td>406</td>
<td>12</td>
<td>50</td>
<td>*</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Cat</td>
<td>63</td>
<td>12</td>
<td>28</td>
<td>30</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Cheetah</td>
<td></td>
<td>*</td>
<td>*</td>
<td>14</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>Chimpanzee</td>
<td>230</td>
<td>20</td>
<td>53</td>
<td>*</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Chipmunk</td>
<td>31</td>
<td>6</td>
<td>8</td>
<td>*</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Cow</td>
<td>284</td>
<td>15</td>
<td>30</td>
<td>*</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Deer</td>
<td>201</td>
<td>8</td>
<td>20</td>
<td>30</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Dog</td>
<td>61</td>
<td>12</td>
<td>20</td>
<td>39</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Donkey</td>
<td>365</td>
<td>12</td>
<td>47</td>
<td>40</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Elephant</td>
<td>660</td>
<td>35</td>
<td>70</td>
<td>25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Elk</td>
<td>250</td>
<td>15</td>
<td>27</td>
<td>45</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Fox</td>
<td>52</td>
<td>7</td>
<td>14</td>
<td>42</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Giraffe</td>
<td>425</td>
<td>10</td>
<td>34</td>
<td>32</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Goat</td>
<td>151</td>
<td>8</td>
<td>18</td>
<td>*</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Gorilla</td>
<td>258</td>
<td>20</td>
<td>54</td>
<td>*</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Guinea pig</td>
<td>68</td>
<td>4</td>
<td>8</td>
<td>*</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hippopotamus</td>
<td>238</td>
<td>41</td>
<td>54</td>
<td>20</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Horse</td>
<td>330</td>
<td>20</td>
<td>50</td>
<td>48</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Kangaroo</td>
<td>36</td>
<td>7</td>
<td>24</td>
<td>40</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Leopard</td>
<td>98</td>
<td>12</td>
<td>23</td>
<td>*</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Lion</td>
<td>100</td>
<td>15</td>
<td>30</td>
<td>50</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Monkey</td>
<td>166</td>
<td>15</td>
<td>37</td>
<td>*</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Moose</td>
<td>240</td>
<td>12</td>
<td>27</td>
<td>*</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Mouse</td>
<td>21</td>
<td>3</td>
<td>4</td>
<td>*</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Opossum</td>
<td>13</td>
<td>1</td>
<td>5</td>
<td>*</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Pig</td>
<td>112</td>
<td>10</td>
<td>27</td>
<td>11</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Puma</td>
<td>90</td>
<td>12</td>
<td>20</td>
<td>*</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Rabbit</td>
<td>31</td>
<td>5</td>
<td>13</td>
<td>35</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Rhinoceros</td>
<td>450</td>
<td>15</td>
<td>45</td>
<td>*</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Sea lion</td>
<td>350</td>
<td>12</td>
<td>30</td>
<td>*</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Sheep</td>
<td>154</td>
<td>12</td>
<td>20</td>
<td>*</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Squirrel</td>
<td>44</td>
<td>10</td>
<td>23</td>
<td>12</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Tiger</td>
<td>105</td>
<td>16</td>
<td>26</td>
<td>*</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Wolf</td>
<td>63</td>
<td>5</td>
<td>13</td>
<td>*</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Zebra</td>
<td>365</td>
<td>15</td>
<td>50</td>
<td>40</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Counts of how many and measurements of how much are called **quantitative variables**. Speed is a quantitative variable because speed is measured on a numerical scale. A variable that groups cases into categories is called a **categorical variable**. Predictor is a categorical variable because it groups the mammals into those who eat other animals and those who don’t. Although the categories are coded 1 if a mammal preys on other animals and 0 if it does not, these numbers just indicate the appropriate category and are not meant to be quantitative. In Display 2.24, the asterisks (*) mark missing values.

### Cases and Variables, Quantitative and Categorical

D6. Classify each variable in Display 2.24 as quantitative or categorical.

### More About Dot Plots

You’ve already seen dot plots, beginning in Chapter 1. As the name suggests, dot plots show individual cases as dots (or other plotting symbols, such as x). When you read a dot plot, keep in mind that different statistical software packages make dot plots in different ways. Sometimes one dot represents two or more cases, and sometimes values have been rounded. With a small data set, different rounding rules can give different shapes.

Display 2.25 shows a dot plot of the speeds of the mammals from Display 2.24. The gap between the cheetah’s speed and that of the other mammals shows up clearly in the dot plot but not in the list of speeds in Display 2.24. Discoveries like this demonstrate why you should always plot your data.

![Dot plot of the speeds of mammals.](image)

As you saw in Section 2.1, a dot plot shows shape, center, and spread. Dot plots tend to work best when
- you have a relatively small number of values to plot
- you want to see individual values, at least approximately
- you want to see the shape of the distribution
- you have one group or a small number of groups you want to compare

### Histograms

A dot plot shows individual cases as dots above a number line. To make a **histogram**, you divide the number line into intervals, called **bins**, and over each bin construct a bar that has a height equal to the number of cases in that bin. In fact, you can think of a histogram as a dot plot with bars drawn around
the dots and the dots erased. The height of the bar becomes a visual substitute for the number of dots. The plot in Display 2.26 is a histogram of the mammal speeds. Like the dot plot of a distribution, a histogram shows shape, center, and spread. The vertical axis gives the number of cases (the frequency or count) represented by each bar. For example, four mammals have speeds of from 30 mi/h up to 35 mi/h.

![Histogram of mammal speeds.](image1)

**Display 2.26** Histogram of mammal speeds.

Most calculators and statistical software packages place a value that falls at the dividing line between two bars into the bar to the right. For example, in Display 2.26, the bar going from 30 to 35 contains cases for which $30 \leq \text{speed} < 35$.

Changing the width of the bars in your histogram can sometimes change your impression of the shape of the distribution. For example, the histogram of the speeds of mammals in Display 2.27 has fewer and wider bars than the histogram in Display 2.26. It shows a more symmetric, bell-shaped distribution, and there appears to be one peak rather than two. There is no “right answer” to the question of which bar width is best, just as there is no rule that tells a photographer when to use a zoom lens for a close-up. Different versions of a picture bring out different features. The job of a data analyst is to find a plot that shows important features of the distribution.

![Histogram of mammal speeds.](image2)

**Display 2.27** Speeds of mammals using a histogram with wider bars.
You can use your calculator to quickly display histograms with different bar widths. [See Calculator Note 2C.] Shown here are the mammal speed data. The numbers below the calculator screens indicate the window settings (minimum $x$, maximum $x$, $x$-scale, minimum $y$, maximum $y$, $y$-scale).

Histograms work best when
• you have a large number of values to plot
• you don’t need to see individual values exactly
• you want to see the general shape of the distribution
• you have only one distribution or a small number of distributions you want to compare
• you can use a calculator or computer to make the plots for you

A histogram shows frequencies on the vertical axis. To change a histogram into a relative frequency histogram, divide the frequency for each bar by the total number of values in the data set and show these relative frequencies on the vertical axis.

**Example: Converting Frequencies to Relative Frequencies**

Four of the 18 mammals have speeds from 30 mi/h up to 35 mi/h. Convert the frequency 4 to a relative frequency.

**Solution**

Four out of 18 is \( \frac{4}{18} \), or approximately 0.22. So about 0.22 of the mammals have speeds in this range.

**Example: Relative Frequency of Life Expectancies**

Display 2.28 shows the relative frequency distribution of life expectancies for 203 countries around the world. How many countries have a life expectancy of at least 70 but less than 75 years? What proportion of the countries have a life expectancy of 70 years or more?
Display 2.28  Life expectancies of people in countries around the world. [Source: Population Reference Bureau, World Population Data Sheet, 2005.]

Solution
The bar including 70 years and up to 75 years has a relative frequency of about 0.30, so the number of countries with a life expectancy of at least 70 years but less than 75 years is about \(0.30 \times 203\), or about 61.

The proportion of countries with a life expectancy of 70 years or greater is the sum of the heights of the three bars to the right of 70—about 0.30 + 0.19 + 0.07, or 0.56.

DISCUSSION

Histograms

D7. In what sense does a histogram with narrow bars, as in Display 2.26, give you more information than a histogram with wider bars, as in Display 2.27? In light of your answer, why don't we always make histograms with very narrow bars?

D8. Does using relative frequencies change the shape of a histogram? What information is lost and gained by using a relative frequency histogram rather than a frequency histogram?

Stemplots

The plot in Display 2.29 is a stem-and-leaf plot, or stemplot, of the mammal speeds. It shows the key features of the distribution and preserves all the original numbers.

Display 2.29  Stemplot of mammal speeds.
In Display 2.29, the numbers on the left, called the stems, are the tens digits of the speeds. The numbers on the right, called the leaves, are the ones digits of the speeds. The leaf for the speed 39 mi/h is printed in bold. If you turn your book 90° counterclockwise, you will see that a stemplot looks something like a dot plot or histogram; you can see the shape, center, and spread of the distribution.

The stemplot in Display 2.30 displays the same information, but with **split stems**: Each stem from the original plot has become two stems. If the ones digit is 0, 1, 2, 3, or 4, it is placed on the first line for that stem. If the ones digit is 5, 6, 7, 8, or 9, it is placed on the second line for that stem.

```
  1   |  1 2
      .
  2   |  0
      .  5
  3   |  0 0 0 2
      .  5 9
  4   |  0 0 0 2
      .  5 8
  5   |  0
  6   |  0
  7   |  0
      3  |  9 represents 39 mi/h
```

**Display 2.30** Stemplot of mammal speeds, using split stems.

Spreading out the stems in this way is similar to changing the width of the bars in a histogram. The goal here, as always, is to find a plot that conveys the essential pattern of the distribution as clearly as possible.

You have compared two data distributions by constructing dot plots on the same scale (see, for example, Display 2.10). Another way to compare two distributions is to construct a back-to-back stemplot. Such a plot for the speeds of predators and nonpredators is shown in Display 2.31. The predators tend to have the faster speeds—or, at least, there are no slow predators!

```
Predator | Nonpredator
---------|-----------
  1 1 2  |  1 2
      .  |  .
  2 0   |  0
      .  5
  3 0 0 2 |  0 0 0 2
      .  5 9
  4 0 0 2 |  0 0 0 2
      .  5 8
  5 0   |  0
  6     |  0
  7     |  0
      3  |  9 represents 39 mi/h
```

**Display 2.31** Back-to-back stemplot of mammal speeds for predators and nonpredators.
Usually, only two digits are plotted on a stemplot, one digit for the stem and one digit for the leaf. If the values contain more than two digits, the values may be either truncated (the extra digits simply cut off) or rounded. For example, if the speeds had been given to the nearest tenth of a mile, 32.6 mi/h could be either truncated to 32 mi/h or rounded to 33 mi/h.

As with the other types of plots, the rules for making stemplots are flexible. Do what seems to work best to reveal the important features of the data.

The stemplot of mammal speeds in Display 2.32 was made by statistical software. Although it looks a bit different from the handmade plot in Display 2.30, it is essentially the same. In the first two lines, \( N = 18 \) means that 18 cases were plotted; \( N^* = 21 \) means that there were 21 cases in the original data set for which speeds were missing; and Leaf Unit = 1.0 means that the ones digits were graphed as the leaves. The numbers in the left column keep track of the number of cases, counting in from the extremes. The 2 on the left in the first line means that there are two cases on that stem. If you skip down three lines, the 4 on the left means that there are a total of four cases on the first four stems (11, 12, 20, and 25).

### Display 2.32

<table>
<thead>
<tr>
<th>Stem-and-leaf of Speeds</th>
<th>N = 18</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leaf Unit = 1.0</td>
<td>N* = 21</td>
</tr>
<tr>
<td>2</td>
<td>1 12</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2 0</td>
</tr>
<tr>
<td>4</td>
<td>2 5</td>
</tr>
<tr>
<td>8</td>
<td>3 0002</td>
</tr>
<tr>
<td>(2)</td>
<td>3 59</td>
</tr>
<tr>
<td>8</td>
<td>4 0002</td>
</tr>
<tr>
<td>4</td>
<td>4 58</td>
</tr>
<tr>
<td>2</td>
<td>5 0</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>7 0</td>
</tr>
</tbody>
</table>

Stemplots are useful when
- you are plotting a single quantitative variable
- you have a relatively small number of values to plot
- you would like to see individual values exactly, or, when the values contain more than two digits, you would like to see approximate individual values
- you want to see the shape of the distribution clearly
- you have two (or sometimes more) groups you want to compare

**DISCUSSION**

**Stemplots**

D9. What information is given by the numbers in the bottom half of the far left column of the plot in Display 2.32? What does the 2 in parentheses indicate?
D10. How might you construct a stemplot of the data on gestation periods for the mammals listed in Display 2.24? Construct the stemplot and describe the shape of the distribution.

**ACTIVITY 2.2a**

**Do Units of Measurement Affect Your Estimates?**

In this activity, you will see whether you and your class estimate lengths better in feet or in meters.

1. Your instructor will split the class randomly into two groups.
2. If you are in the first group, you will estimate the length of your classroom in feet. If you are in the second group, you will estimate the length of the classroom in meters (without estimating first in feet and then converting to meters). Do this by looking at the length of the room; no pacing the length of the room is allowed.
3. Find an appropriate way to plot the two data sets so that you can compare their shapes, centers, and spreads.
4. Do the students in your class tend to estimate more accurately in feet or in meters? What is the basis for your decision?
5. Why split the class randomly into two groups instead of simply letting the left half of the room estimate in feet and the right half estimate in meters?

**Bar Charts for Categorical Data**

You now have three different types of plots to use with quantitative variables. What about categorical variables? You could make a dot plot, or you could make what looks like a histogram but is called a bar chart or bar graph. There is one bar for each category, and the height of the bar tells the frequency. A bar chart has categories on the horizontal axis, whereas a histogram has measurements—values from a quantitative variable.

The bar chart in Display 2.33 shows the frequency of mammals that fall into the categories “wild” and “domesticated,” coded 1 and 0, respectively. (Note that the bars are separated so that there is no suggestion that the variable can take on a value of, say, 0.5.)

![Display 2.33 Bar chart showing frequency of domesticated (0) and wild (1) mammals.](image-url)
The relative frequency bar chart in Display 2.34 shows the proportion of the female labor force age 25 and older in the United States who fall into various educational categories. The coding used in the display is as follows:

1: none—8th grade  
2: 9th grade—11th grade  
3: high school graduate  
4: some college, no degree  
5: associate degree  
6: bachelor’s degree  
7: master’s degree  
8: professional degree  
9: doctorate degree

The educational categories in Display 2.34 have a natural order from least education to most education and are coded with the numbers 1 through 9. Note that if you compute the mean of this distribution, there is no reasonable way to interpret it. However, it does make sense to summarize this distribution using the mode: More women fall into the category “high school graduate” than into any other category. Thus, the numbers 1 through 9 are best thought of as representing an ordered categorical variable, not a quantitative variable.

You will learn more about the analysis of categorical data in Chapter 10.

**Bar Charts**

D11. In the bar chart in Display 2.33, would it matter if the order of the bars were reversed? In the bar chart in Display 2.34, would it matter if the order of the first two bars were reversed? Comment on how we might define two different types of categorical variables.

**Summary 2.2: Graphical Displays of Distributions**

When a variable is quantitative, you can use dot plots, stemplots (stem-and-leaf plots), and histograms to display the distribution of values. From each plot, you can see the shape, center, and spread of the distribution. However, the amount of
detail varies, and you should choose a plot that fits both your data set and your reason for analyzing it.

- Stemplots can retain the actual data values.
- Dot plots are best used with a small number of values and show roughly where the values lie on a number line.
- Histograms show only intervals of values, losing the actual data values, and are most appropriate for large data sets.

A bar chart shows the distribution of a categorical variable.

When you look at a plot, you should attempt to answer these questions:

- Where did this data set come from?
- What are the cases and the variables?
- What are the shape, center, and spread of this distribution? Does the distribution have any unusual characteristics, such as clusters, gaps, or outliers?
- What are possible interpretations or explanations of the patterns you see in the distribution?

### Practice

#### More About Dot Plots

**P6.** In the listing of the Westvaco data in Display 1.1 on page 5, which variables are quantitative? Which are categorical?

**P7.** Select a reasonable scale, and make a dot plot of the gestation periods of the mammals listed in Display 2.24 on page 43. Write a sentence using shape, center, and spread to summarize the distribution of gestation periods for the mammals. What kinds of mammals have longer gestation periods?

#### Histograms

**P8.** Make histograms of the average longevities and the maximum longevities from Display 2.24. Describe how the distributions differ in terms of shape, center, and spread. Why do these differences occur?

**P9.** Convert your histograms from P8 of the average longevities and maximum longevities of the mammals to relative frequency histograms. Do the shapes of the histograms change?

**P10.** Using the relative frequency histogram of life expectancy in countries around the world (Display 2.28 on page 47), estimate the proportion of countries with a life expectancy of less than 50 years. Then estimate the number of countries with a life expectancy of less than 50 years. Describe the shape, center, and spread of this distribution.

#### Stemplots

**P11.** Make a back-to-back stemplot of the average longevities and maximum longevities from Display 2.24 on page 43. Compare the two distributions.
Bar Charts for Categorical Data

P12. Display 2.35, educational attainment of the male labor force, is the counterpart of Display 2.34. What are the cases, and what is the variable? Describe the distribution you see here. How does the distribution of female education compare to the distribution of male education? Why is it better to look at relative frequency bar charts rather than frequency bar charts to make this comparison?

Display 2.35  The male labor force age 25 years and older by educational attainment.

P13. Using the Westvaco data in Display 1.1 on page 5, make a bar chart showing the number of workers laid off in each round. In addition to a bar showing layoffs, for each of the five rounds, include a bar showing the number of workers not laid off. Then make a relative frequency bar chart. Describe any patterns you see.

Exercises

E15. The dot plot in Display 2.36 shows the distribution of the ages of pennies in a sample collected by a statistics class.

Display 2.36  Age of pennies. Each dot represents four points.

E16. Suppose you collect this information for each student in your class: age, hair color, number of siblings, gender, and miles he or she lives from school. What are the cases? What are the variables? Classify each variable as quantitative or categorical.
E17. Using your knowledge of the variables and what you think the shape of the distribution might be, match each variable in this list with the appropriate histogram in Display 2.37.

I. scores on a fairly easy examination in statistics
II. heights of a group of mothers and their 12-year-old daughters
III. numbers of medals won by medal-winning countries in the 2004 Summer Olympics
IV. weights of grown hens in a barnyard

A.  

B.  

C.  

D.  

Display 2.37 Four histograms with different shapes.

E18. Using the technology available to you, make histograms of the average longevity and maximum longevity data in Display 2.24 on page 43, using bar widths of 4, 8, and 16 years. Comment on the main features of the shapes of these distributions and determine which bar width appears to display these features best.

E19. Rewrite each sentence so that it states a relative frequency rather than a count.

a. Six students in a class of 30 got an A.

b. Out of the 50,732 people at a concert, 24,021 bought a T-shirt.

E20. Convert the histogram in Display 2.38 into a relative frequency histogram.

Display 2.38 Ages of 1000 people.

E21. Display 2.39 shows the distribution of the heights of U.S. males between the ages of 18 and 24. The heights are rounded to the nearest inch.


a. Draw a smooth curve to approximate the histogram.

b. Without doing any computing, estimate the mean and standard deviation.

c. Estimate the proportion of men age 18 to 24 who are 74 in. tall or less.

d. Estimate the proportion of heights that fall below 68 in.

e. Why should you say that the distribution of heights is “approximately” normal rather than simply saying that it is normally distributed?

a. Without doing any computing, estimate the mean and standard deviation.

b. Roughly what percentage of the SAT I math scores would you estimate are within one standard deviation of the mean?

c. For SAT I critical reading scores, the shape was similar, but the mean was 10 points lower and the standard deviation was 2 points smaller. Draw a smooth curve to show the distribution of SAT I critical reading scores.


E23. In this section, you looked at various characteristics of mammals.

a. Would you predict that wild mammals or domesticated mammals generally have greater longevity?

b. Using the data in Display 2.24 on page 43, make a back-to-back stemplot to compare the average longevities.

c. Write a short summary comparing the two distributions.


E24. The plots in Display 2.41 show a form of back-to-back histogram called a population pyramid. Describe how the population distribution of the United States differs from the population distribution of Mexico.
Measures of Center and Spread

So far you have relied on visual methods for estimating summary statistics to measure center and spread. In this section, you will learn how to compute exact values of those same summary statistics directly from the data.

Measures of Center

The two most commonly used measures of center are the mean and the median.

The mean, \( \bar{x} \), is the same number that many people call the “average.” To compute the mean, sum all the values of \( x \) and divide by the number of values, \( n \):

\[
\bar{x} = \frac{\sum x}{n}
\]

(The symbol \( \sum \), for sum, means to add up all the values of \( x \).)
The mean is the balance point of a distribution. To estimate the mean visually on a dot plot or histogram, find where you would have to place a finger below the horizontal axis in order to balance the distribution, as if it were a tray of blocks (see Display 2.43).

The median is the value that divides the data into halves, as shown in Display 2.44. To find it for an odd number of values, list all the values in order and select the middle one. If there are $n$ values and $n$ is odd, you will find the median at position $\frac{n + 1}{2}$. If $n$ is even, the median is the average of the values on either side of position $\frac{n + 1}{2}$.

**Example: Effect of Round 2 Layoffs on Measures of Center**

The ages of the hourly workers involved in Round 2 of the layoffs at Westvaco were 25, 33, 35, 38, 48, 55, 55*, 55*, 56, and 64* (* indicates laid off in Round 2). The two dot plots in Display 2.45 show the distributions of hourly workers before and after the second round of layoffs. What was the effect of Round 2 on the mean age? On the median age?
Solution

Means

Before: The sum of the ten ages is 464, so the mean age is \( \frac{464}{10} \), or 46.4 years.

After: There are seven ages and their sum is 290, so the mean age is \( \frac{290}{7} \), or 41.4 years.

The layoffs reduced the mean age by 5 years.

Medians

Before: Because there are ten ages, \( n = 10 \), so \( \frac{(n + 1)}{2} = \frac{(10 + 1)}{2} \) or 5.5, and the median is halfway between the fifth ordered value, 48, and the sixth ordered value, 55. The median is \( \frac{(48 + 55)}{2} \), or 51.5 years.

After: There are seven ages, so \( \frac{(n + 1)}{2} = \frac{(7 + 1)}{2} \) or 4. The median is the fourth ordered value, or 38 years.

The layoffs reduced the median age by 13.5 years.

DISCUSSION

Measures of Center

D12. Find the mean and median of each ordered list, and contrast their behavior.
   a. 1, 2, 3
   b. 1, 2, 6
   c. 1, 2, 9
   d. 1, 2, 297

D13. As you saw in D12, typically the mean is affected more than the median by an outlier.
   a. Use the fact that the median is the halfway point and the mean is the balance point to explain why this is true.
   b. For the distributions of mammal speeds in Display 2.31 on page 48, the means are 43.5 mi/h for predators and 31.5 mi/h for nonpredators. The medians are 40.5 mi/h and 33.5 mi/h, respectively. What about the distributions causes the means to be farther apart than the medians?
   c. What about the shapes of the plots in Display 2.45 explains why the means change so much less than the medians?
Measuring Spread Around the Median: Quartiles and IQR

You can locate the median of a distribution by dividing your data into a lower and upper half. You can use the same idea to measure spread: Find the values that divide each half in half again. These two values, the lower quartile, \( Q_1 \), and the upper quartile, \( Q_3 \), together with the median, divide your data into four quarters. The distance between the upper and lower quartiles, called the interquartile range, or IQR, is a measure of spread.

\[ IQR = Q_3 - Q_1 \]

San Francisco, California, and Springfield, Missouri, have about the same median temperature over the year. In San Francisco, half the months of the year have a normal temperature above 56.5°F, half below. In Springfield, half the months have a normal temperature above 57°F, half below. If you judge by these medians, the difference hardly matters. But if you visit San Francisco, you better take a jacket, no matter what month you go. If you visit Springfield, take your shorts and a T-shirt in the summer and a heavy coat in the winter. The difference in temperatures between the two cities is not in their centers but in their variability. In San Francisco, the middle 50% of normal monthly temperatures lie in a narrow 9° interval between 52.5°F and 61.5°F, whereas in Springfield the middle 50% of normal monthly temperatures range over a 31° interval, varying from 40.5°F to 71.5°F. In other words, the IQR is 9°F for San Francisco and 31°F for Springfield.

Finding the Quartiles and IQR

If you have an even number of cases, finding the quartiles is straightforward: Order your observations, divide them into a lower and upper half, and then divide each half in half. If you have an odd number of cases, the idea is the same, but there’s a question of what to do with the middle value when you form the upper and lower halves.

There is no one standard answer. Different statistical software packages use different procedures that can give slightly different values for the quartiles. In this book, the procedure is to omit the middle value when you form the two halves.

Example: Finding the Quartiles and IQR for Workers’ Ages

Find the quartiles and IQR for the ages of the hourly workers at Westvaco before and after Round 2 of the layoffs.

Solution

Before: There are ten ages: 25, 33, 35, 38, 48, 55, 55, 55, 56, 64. Because \( n \) is even, the median is halfway between the two middle values, 48 and 55, so it is 51.5. The lower half of the data is made up of the first five ordered values, and the median of the lower half is the third value, so \( Q_1 \) is 35. The upper half of the data is the set of the five largest values, and the median of these is again the third value, so \( Q_3 \) is 55. The IQR is 55 − 35, or 20.
After: After the three workers are laid off in Round 2, there are seven ages: 25, 33, 35, 38, 48, 55, 56. Because \( n \) is odd, the median is the middle value, 38. Omit this one number. The lower half of the data is made up of the three ordered values to the left of position 4. The median of these is the second value, so \( Q_1 \) is 33. The upper half of the data is the set of the three ordered values to the right of position 4, and the median of these is again the second value, so \( Q_3 \) is 55. The IQR is 55 – 33, or 22.

**Finding the Quartiles and IQR**

D14. Here are the medians and quartiles for the speeds of the domesticated and wild mammals:

<table>
<thead>
<tr>
<th></th>
<th>( Q_1 )</th>
<th>Median</th>
<th>( Q_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domesticated</td>
<td>30</td>
<td>37</td>
<td>40</td>
</tr>
<tr>
<td>Wild</td>
<td>27.5</td>
<td>36</td>
<td>43.5</td>
</tr>
</tbody>
</table>

a. Use the information in Display 2.24 on page 43 to verify these values, and then use them to summarize and compare the two distributions.

b. Why might the speeds of domesticated mammals be less spread out than the speeds of wild mammals?

D15. The following quote is from the mystery *The List of Adrian Messenger*, by Philip MacDonald (Garden City, NY: Doubleday, 1959, p. 188). Detective Firth asks Detective Seymour if eyewitness accounts have provided a description of the murderer:

“Descriptions?” he said. “You must’ve collected quite a few. How did they boil down?”

“To a no-good norm, sir,” Seymour shrugged wearily. “They varied so much, the average was useless.”

Explain what Detective Seymour means.

**Five-Number Summaries, Outliers, and Boxplots**

The visual, verbal, and numerical summaries you’ve seen so far tell you about the middle of a distribution but not about the extremes. If you include the minimum and maximum values along with the median and quartiles, you get the five-number summary.
The five-number summary for a set of values:

- **Minimum**: the smallest value
- **Lower or first quartile, Q₁**: the median of the lower half of the ordered set of values
- **Median or second quartile**: the value that divides the ordered set of values into halves
- **Upper or third quartile, Q₃**: the median of the upper half of the ordered set of values
- **Maximum**: the largest value

The difference of the maximum and the minimum is called the **range**.

Display 2.46 shows the five-number summary for the speeds of the mammals listed in Display 2.24.

| 1 | 12 | 2 | 05 | 3 | 00259 | 4 | 00258 | 5 | 0 | 6 | 7 | 0 |
|---|----|---|----|---|-------|---|-------|---|---|--|--|--|--|
| **min** | 11 | **Q₁** | 30 | **median** | 37 | **Q₃** | 42 | **max** | 70 |

**Display 2.46** Five-number summary for the mammal speeds.

Display 2.47 shows a boxplot of the mammal speeds. A boxplot (or box-and-whiskers plot) is a graphical display of the five-number summary. The “box” extends from Q₁ to Q₃, with a line at the median. The “whiskers” run from the quartiles to the extreme values.

**Display 2.47** Boxplot of mammal speeds.

The maximum speed of 70 mi/h for the cheetah is 20 mi/h from the next fastest mammal (the lion) and 28 mi/h from the nearest quartile. It is handy to have a version of the boxplot that shows isolated cases—outliers—such as the cheetah. Informally, outliers are any values that stand apart from the rest. This rule often is used to identify outliers.

A value is an **outlier** if it is more than 1.5 times the IQR from the nearest quartile.

1.5 • IQR rule for outliers

Note that “more than 1.5 times the IQR from the nearest quartile” is another way of saying “either greater than Q₃ plus 1.5 times IQR or less than Q₁ minus 1.5 times IQR.”
**Example: Outliers in the Mammal Speeds**

Use the $1.5 \cdot IQR$ rule to identify outliers and the largest and smallest non-outliers among the mammal speeds.

**Solution**

From Display 2.46, $Q_1 = 30$ and $Q_3 = 42$, so the $IQR$ is $42 - 30 = 12$, and $1.5 \cdot IQR$ equals 18.

*At the low end:*

$$Q_1 - 1.5 \cdot IQR = 30 - 18 = 12$$

The pig, at 11 mi/h, is an outlier.
The squirrel, at 12 mi/h, is the smallest non-outlier.

*At the high end:*

$$Q_3 + 1.5 \cdot IQR = 42 + 18 = 60$$

The cheetah, at 70 mi/h, is an outlier.
The lion, at 50 mi/h, is the largest non-outlier.

A modified boxplot, shown in Display 2.48, is like the basic boxplot except that the whiskers extend only as far as the largest and smallest non-outliers (sometimes called *adjacent values*) and any outliers appear as individual dots or other symbols.

![Modified Boxplot](image)

Display 2.48  Modified boxplot of mammal speeds.

Boxplots are particularly useful for comparing several distributions.

**Example: Boxplots That Show Outliers**

Display 2.49 shows side-by-side modified boxplots of average longevity for wild and domesticated mammals. Compare the two distributions.

![Boxplots](image)

Display 2.49  Comparison of average longevity.
Solution

The boxplot for domesticated animals has no median line. So many domesticated animals had an average longevity of 12 years that it is both the median and the upper quartile. These plots show that species of domestic mammals typically have median average longevities of about 12 years, with about the middle half of these average longevities falling between 8 and 12 years. The average longevity of wild mammals centers at about the same place, but the wild mammal average longevities have more variability, with the middle half between about 7.5 and 15.5 years. Both shapes are roughly symmetric except for some unusually large average longevities—two wild mammals have average longevities of more than 30 years.

[See Calculator Note 2D to learn how to display regular and modified boxplots and five-number summaries on your calculator.]

Boxplots are useful when you are plotting a single quantitative variable and
• you want to compare the shapes, centers, and spreads of two or more distributions
• you don't need to see individual values, even approximately
• you don't need to see more than the five-number summary but would like outliers to be clearly indicated

Five-Number Summaries, Outliers, and Boxplots

D16. Test your ability to interpret boxplots by answering these questions.

a. Approximately what percentage of the values in a data set lie within the box? Within the lower whisker, if there are no outliers? Within the upper whisker, if there are no outliers?

b. How would a boxplot look for a data set that is skewed right? Skewed left? Symmetric?

c. How can you estimate the IQR directly from a boxplot? How can you estimate the range?

d. Is it possible for a boxplot to be missing a whisker? If so, give an example. If not, explain why not.

e. Contrast the information you can learn from a boxplot with what you can learn from a histogram. List the advantages and disadvantages of each type of plot.
Measuring Spread Around the Mean: The Standard Deviation

There are various ways you can measure the spread of a distribution around its mean. Activity 2.3a gives you a chance to create a measure of your own.

**ACTIVITY 2.3a**

**Comparing Hand Spans: How Far Are You from the Mean?**

**What you’ll need:** a ruler

1. Spread your hand on a ruler and measure your hand span (the distance from the tip of your thumb to the tip of your little finger when you spread your fingers) to the nearest half centimeter.
2. Find the mean hand span for your group.
3. Make a dot plot of the results for your group. Write names or initials above the dots to identify the cases. Mark the mean with a wedge (▲) below the number line.
4. Give two sources of variability in the measurements. That is, give two reasons why all the measurements aren’t the same.
5. How far is your hand span from the mean hand span of your group? How far from the mean are the hand spans of the others in your group?
6. Make a plot of differences from the mean. Again label the dots with names or initials. What is the mean of these differences? Tell how to get the second plot from the first without computing any differences.
7. Using the idea of differences from the mean, invent at least two measures that give a “typical” distance from the mean.
8. Compare your measures with those of the other groups in your class. Discuss the advantages and disadvantages of each group’s method.

The differences from the mean, \(x - \bar{x}\), are called deviations. The mean is the balance point of the distribution, so the set of deviations from the mean will always sum to zero.

**Deviations from the mean sum to zero:**

\[ \sum(x - \bar{x}) = 0 \]

**Example: Deviations from the Mean**

Find the deviations from the mean for the predators’ speeds and verify that the sum of these deviations is 0. Which predator’s speed is farthest from the mean?
Solution
The speeds 30, 30, 39, 42, 50, and 70 mi/h have mean 43.5 mi/h. The deviations from the mean are

\[
\begin{align*}
30 - 43.5 &= -13.5 \\
30 - 43.5 &= -13.5 \\
39 - 43.5 &= -4.5 \\
42 - 43.5 &= -1.5 \\
50 - 43.5 &= 6.5 \\
70 - 43.5 &= 26.5
\end{align*}
\]

The sum of the deviations is \(-13.5 + (-13.5) + (-4.5) + (-1.5) + 6.5 + 26.5\), which equals 0. The cheetah’s speed, 70 mi/h, is farthest from the mean.

How can you use the deviations from the mean to get a measure of spread? You can’t simply find the average of the deviations, because you will get 0 every time. As you might have suggested in the activity, you could find the average of the absolute values of the deviations. That gives a perfectly reasonable measure of spread, but it does not turn out to be very easy to use or very useful. Think of how hard it is to deal with an equation that has sums of absolute values in it, for example, \(y = |x - 1| + |x - 2| + |x - 3|\). On the other hand, if you square the deviations, which also gets rid of the negative signs, you get a sum of squares. Such a sum is always quadratic no matter how many terms there are, for example, \(y = (x - 1)^2 + (x - 2)^2 + (x - 3)^2 = 3x^2 - 12x + 14\).

The measure of spread that incorporates the square of the deviations is the standard deviation, abbreviated \(SD\) or \(s\), that you met in Section 2.1. Because sums of squares really are easy to work with mathematically, the SD offers important advantages that other measures of spread don’t give you. You will learn more about these advantages in Chapter 7. The formula for the standard deviation, \(s\), is given in the box.

### Formula for the Standard Deviation, \(s\)

\[
s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}
\]

The square of the standard deviation, \(s^2\), is called the variance.

Calculators might label the two versions \(\sigma\) and \(\sigma_{n-1}\), or \(\sigma\) and \(s\).  

It might seem more natural to divide by \(n\) to get the average of the squared deviations. In fact, two versions of the standard deviation formula are used: One divides by the sample size, \(n\); the other divides by \(n - 1\). Dividing by \(n - 1\) gives a slightly larger value. This is useful because otherwise the standard deviation computed from a sample would tend to be smaller than the standard deviation of the population from which the sample came. (You will learn more about this in Chapter 7.) In practice, dividing by \(n - 1\) is almost always used for real data even if they aren’t a sample from a larger population.
**Example: The Standard Deviation of Mammal Longevity**

Compute the standard deviation of the average longevity of domesticated mammals from Display 2.24 on page 43.

**Solution**

The table in Display 2.50 is a good way to organize the steps. First find the mean longevity, $\bar{x}$, and then subtract it from each observed value $x$ to get the deviations, $x - \bar{x}$. Square each deviation to get $(x - \bar{x})^2$.

<table>
<thead>
<tr>
<th>Case</th>
<th>Longevity, $x$</th>
<th>Mean, $\bar{x}$</th>
<th>Deviation, $x - \bar{x}$</th>
<th>Squared Deviation, $(x - \bar{x})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cat</td>
<td>12</td>
<td>11</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Cow</td>
<td>15</td>
<td>11</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>Dog</td>
<td>12</td>
<td>11</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Donkey</td>
<td>12</td>
<td>11</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Goat</td>
<td>8</td>
<td>11</td>
<td>-3</td>
<td>9</td>
</tr>
<tr>
<td>Guinea pig</td>
<td>4</td>
<td>11</td>
<td>-7</td>
<td>49</td>
</tr>
<tr>
<td>Horse</td>
<td>20</td>
<td>11</td>
<td>9</td>
<td>81</td>
</tr>
<tr>
<td>Pig</td>
<td>10</td>
<td>11</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>Rabbit</td>
<td>5</td>
<td>11</td>
<td>-6</td>
<td>36</td>
</tr>
<tr>
<td>Sheep</td>
<td>12</td>
<td>11</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>110</td>
<td>110</td>
<td>0</td>
<td>196</td>
</tr>
</tbody>
</table>

**Display 2.50**  Steps in computing the standard deviation.

To compute the standard deviation, sum the squared deviations, divide the sum by $n - 1$, and finally take the square root:

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} = \sqrt{\frac{196}{10 - 1}} \approx 4.67$$

[You can organize the steps of calculating the standard deviation on your calculator. See Calculator Note 2E.]

Your calculator will compute the summary statistics for a set of data. [See Calculator Note 2F.] Here are the summary statistics for the domesticated mammal longevity data. Note that the standard deviation calculated in the previous example is denoted as $S_x$. Note also that the five-number summary is shown.
The Standard Deviation

D17. Refer to the previous example for mammal longevities.
   a. Does 4.67 years seem like a typical distance from the mean of 11 years for the average longevities in the example?
   b. The average longevities are measured in years. What is the unit of measurement for the mean? For the standard deviation? For the variance? For the interquartile range? For the median?

D18. The standard deviation, if you look at it the right way, is a generalization of the usual formula for the distance between two points. How does the formula for the standard deviation remind you of the formula for the distance between two points?

D19. What effect does dividing by \( n - 1 \) rather than by \( n \) have on the standard deviation? Does which one you divide by matter more with a large number of values or with a small number of values?

Summaries from a Frequency Table

To find the mean of the numbers 5, 5, 5, 5, 5, 5, 8, 8, and 8, you could sum them and divide their sum by how many numbers there are. However, you could get the same answer faster by taking advantage of the repetitions:

\[
\bar{x} = \frac{5 \cdot 6 + 8 \cdot 3}{6 + 3} = \frac{30 + 24}{9} = 6
\]

You can use formulas to find the mean and standard deviation of values in a frequency table, like the one in Display 2.51 in the example on the next page.

Formulas for the Mean and Standard Deviation of Values in a Frequency Table

If each value \( x \) occurs with frequency \( f \), the mean of a frequency table is given by

\[
\bar{x} = \frac{\sum x \cdot f}{n}
\]

The standard deviation is given by

\[
s = \sqrt{\frac{\sum (x - \bar{x})^2 \cdot f}{n - 1}}
\]

where \( n \) is the sum of the frequencies, or \( n = \sum f \).
Example: Mean and Standard Deviation of Coin Values

Suppose you have five pennies, three nickels, and two dimes. Find the mean value of the coins and the standard deviation.

Solution

The table in Display 2.51 shows a way to organize the steps in computing the mean using the formula for the mean of values in a frequency table.

<table>
<thead>
<tr>
<th>Value</th>
<th>Frequency</th>
<th>$x \cdot f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Penny</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Nickel</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Dime</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>Sum</td>
<td>10</td>
<td>40</td>
</tr>
</tbody>
</table>

$$
\bar{x} = \frac{\sum x \cdot f}{n} = \frac{40}{10} = 4
$$

Display 2.51 Steps in computing the mean of a frequency table.

Display 2.52 gives an extended version of the table, designed to organize the steps in computing both the mean and the standard deviation.

<table>
<thead>
<tr>
<th>Value</th>
<th>Frequency</th>
<th>$x \cdot f$</th>
<th>$x - \bar{x}$</th>
<th>$(x - \bar{x})^2$</th>
<th>$(x - \bar{x})^2 \cdot f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Penny</td>
<td>1</td>
<td>5</td>
<td>-3</td>
<td>9</td>
<td>45</td>
</tr>
<tr>
<td>Nickel</td>
<td>5</td>
<td>3</td>
<td>15</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Dime</td>
<td>10</td>
<td>2</td>
<td>20</td>
<td>6</td>
<td>36</td>
</tr>
<tr>
<td>Sum</td>
<td>10</td>
<td>40</td>
<td>120</td>
<td></td>
<td>120</td>
</tr>
</tbody>
</table>

$$
s = \sqrt{\frac{\sum (x - \bar{x})^2 \cdot f}{n - 1}} = \sqrt{\frac{120}{9}} \approx 3.65
$$

Display 2.52 Steps in computing the standard deviation of values in a frequency table.

[See Calculator Note 2F to learn how to compute numerical summaries from a frequency table.]

DISCUSSION

Summaries from a Frequency Table

D20. Explain why the formula for the standard deviation in the box on page 67 gives the same result as the formula on page 65.
Summary 2.3: Measures of Center and Spread

Your first step in any data analysis should always be to look at a plot of your data, because the shape of the distribution will help you determine what summary measures to use for center and spread.

- To describe the center of a distribution, the two most common summaries are the median and the mean. The median, or halfway point, of a set of ordered values is either the middle value (if \( n \) is odd) or halfway between the two middle values (if \( n \) is even). The mean, or balance point, is the sum of the values divided by the number of values.
- To measure spread around the median, use the interquartile range, or IQR, which is the width of the middle 50% of the data values and equals the distance from the lower quartile to the upper quartile. The quartiles are the medians of the lower half and upper half of the ordered list of values.
- To measure spread around the mean, use the standard deviation. To compute the standard deviation for a data set of size \( n \), first find the deviations from the mean, then square them, sum the squared deviations, divide by \( n - 1 \), and take the square root.

A boxplot is a useful way to compare the general shape, center, and spread of two or more distributions with a large number of values. A modified boxplot also shows outliers. An outlier is any value more than 1.5 times the IQR from the nearest quartile.

Practice

Measures of Center

P14. Find the mean and median of these ordered lists.
   a. 1 2 3 4 5 b. 1 2 3 4 5
   c. 1 2 3 4 5 6 d. 1 2 3 4 5 . . . 97 98
   e. 1 2 3 4 5 . . . 97 98 99

P15. Five 3rd graders, all about 4 ft tall, are standing together when their teacher, who is 6 ft tall, joins the group. What is the new mean height? The new median height?

P16. The stemplots in Display 2.53 show the life expectancies (in years) for females in the countries of Africa and Europe. The means are 53.6 years for Africa and 79.3 years for Europe.
   a. Find the median life expectancy for each set of countries.
   b. Is the mean or the median smaller for each distribution? Why is this so?

Display 2.53 Female life expectancies in Africa and Europe. [Source: Population Reference Bureau, World Population Data Sheet, 2005.]
Measuring Spread Around the Median: Quartiles and IQR

P17. Find the quartiles and IQR for these ordered lists.
   a. 1 2 3 4 5 6
   b. 1 2 3 4 5 6 7
   c. 1 2 3 4 5 6 7 8
   d. 1 2 3 4 5 6 7 8 9

P18. Display 2.54 shows a back-to-back stemplot of the average longevity of predators and nonpredators.

<table>
<thead>
<tr>
<th>Predators</th>
<th>Nonpredators</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0 3 4</td>
</tr>
<tr>
<td>7 5</td>
<td>5 5 6 7 8 8</td>
</tr>
<tr>
<td>2 2 2 2 2</td>
<td>1 0 0 2 2 2 2</td>
</tr>
<tr>
<td>6 5</td>
<td>5 5 5 5 5 5</td>
</tr>
<tr>
<td>2</td>
<td>0 0 0 0</td>
</tr>
<tr>
<td>5</td>
<td>·</td>
</tr>
<tr>
<td>3</td>
<td>· 5</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

   1 | 5 stands for 15 years

Display 2.54 Average longevities of predators and nonpredators.

a. By counting on the plot, find the median and quartiles for each group of mammals.
b. Write a pair of sentences summarizing and comparing the shape, center, and spread of the two distributions.

Five-Number Summaries, Outliers, and Boxplots

P19. The boxplot in Display 2.55 shows the number of viewers who watched the 101 prime-time network television shows in the week that Seinfeld aired its last new episode.

Display 2.55 Modified boxplot of number of viewers of prime-time television shows.

a. Seinfeld had more viewers than any other show. About how many viewers did it have?

P20. Use the medians and quartiles from D14 on page 60 and the data in Display 2.24 on page 43 to construct side-by-side boxplots of the speeds of wild and domesticated mammals. (Don't show outliers in these plots.)

P21. The stemplot of average mammal longevities is shown in Display 2.56.

Display 2.56 Average longevity (in years) of 38 mammals.

a. Use the stemplot to find the five-number summary.
b. Find the IQR.
c. Compute \( Q_1 - 1.5 \times IQR \). Identify any outliers (give the animal name and longevity) at the low end.
d. Identify an outlier at the high end and the largest non-outlier.
e. Draw a modified boxplot.

The Standard Deviation

P22. Verify that the sum of the deviations from the mean is 0 for the numbers 1, 2, 4, 6, and 9. Find the standard deviation.

P23. Without computing, match each list of numbers in the left column with its standard deviation in the right column. Check any answers you aren't sure of by computing.

a. 1 1 1 1  
   i. 0
b. 1 2 2  
   ii. 0.058
   
c. 1 2 3 4 5  
   iii. 0.577
d. 10 20 20  
   iv. 1.581
e. 0 1 0 2 0.2  
   v. 3.162
f. 0 2 4 6 8  
   vi. 3.606
g. 0 0 0 5 6 6 8 8  
   vii. 5.774
2.3 Measures of Center and Spread

Summaries from a Frequency Table

P24. Display 2.57 shows the data on family size for two representative sets of 100 families, one set from 1967 and the other from 1997.

<table>
<thead>
<tr>
<th>Number of Children</th>
<th>Number of Families</th>
<th>Number of Children</th>
<th>Number of Families</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>1</td>
<td>22</td>
</tr>
<tr>
<td>2</td>
<td>21</td>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>28</td>
<td>3</td>
<td>18</td>
</tr>
<tr>
<td>4</td>
<td>17</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>8</td>
<td>2</td>
</tr>
</tbody>
</table>


a. Try to visualize the shapes of the two distributions. Are they symmetric, skewed left, or skewed right?

b. Find the median number of children per family for 1967.

c. Use the formulas to compute the mean and standard deviation of the 1967 distribution.

P25. Refer to Display 2.57.

a. Use the formulas for the mean and standard deviation of values in a frequency table to compute the mean number of children per family and the standard deviation for the 1997 distribution.

b. Find the median number of children for 1997.

c. What are the positions of the quartiles in an ordered list of 100 numbers? Find the quartiles for the 1967 distribution and compute the \( IQR \). Do the same for the 1997 distribution.

d. Write a comparison of the shape, center, and spread of the distributions for the two years.

Exercises

E29. The mean of a set of seven values is 25. Six of the values are 24, 47, 34, 10, 22, and 28. What is the 7th value?

E30. The sum of a set of values is 84, and the mean is 6. How many values are there?

E31. Three histograms and three boxplots appear in Display 2.58. Which boxplot displays the same information as

a. histogram A?

b. histogram B?

c. histogram C?

A. B. C.

Display 2.58 Match the histograms with their boxplots.
E32. The test scores of 40 students in a first-period class were used to construct the first boxplot in Display 2.59, and the test scores of 40 students in a second-period class were used to construct the second. Can the third plot be a boxplot of the combined scores of the 80 students in the two classes? Why or why not?

Display 2.59 Boxplots of three sets of test scores.

E33. Make side-by-side boxplots of the speeds of predators and nonpredators. (The stemplot in Display 2.31 on page 48 shows the values ordered.) Are the boxplots or the back-to-back stemplot in Display 2.31 better for comparing these speeds? Explain.

E34. The U.S. Supreme Court instituted a temporary ban on capital punishment between 1967 and 1976. Between 1977 and 2000, 31 states carried out 683 executions. (The other 19 states either did not have a death penalty or executed no one.) The five states that executed the most prisoners were Texas (239), Virginia (81), Florida (50), Missouri (46), and Oklahoma (30). The remaining 26 states carried out these numbers of executions: 26, 25, 23, 23, 22, 16, 12, 11, 8, 8, 7, 6, 4, 3, 3, 3, 3, 2, 2, 2, 1, 1, 1. For all 50 states, what was the mean number of executions per state? The median number? What were the quartiles? Draw a boxplot, showing any outliers, of the number of executions for all 50 states. [Source: U.S. Department of Justice, Bulletin: Capital Punishment 2000.]

E35. Make a back-to-back stemplot comparing the ages of those retained and those laid off among the salaried workers in the engineering department at Westvaco (see Display 1.1 on page 5). Find the medians and quartiles, and use them to write a verbal comparison of the two distributions.

E36. The boxplots below show the average longevity of mammals, from Display 2.24.

a. Using only the basic boxplot in Display 2.60, show that there must be at least one outlier in the set of average longevity.

Display 2.60 Boxplot of average longevity of mammals.

b. How many outliers are there in the modified boxplot of average longevity in Display 2.61?

Display 2.61 Modified boxplot of average longevity of mammals, showing outliers.

c. How many outliers are shown in Display 2.49 on page 62? How can that be, considering the boxplot in Display 2.61?

E37. No computing should be necessary to answer these questions.

a. The mean of each of these sets of values is 20, and the range is 40. Which set has the largest standard deviation? Which has the smallest?

I. 0 10 20 30 40
II. 0 0 20 40 40
III. 0 19 20 21 40

b. Two of these sets of values have a standard deviation of about 5. Which two?

I. 5 5 5 5 5
II. 10 10 10 20 20
III. 6 8 10 12 14 16 18 20 22
IV. 5 10 15 20 25 30 35 40 45

E38. The standard deviation of the first set of values listed here is about 32. What is the standard deviation of the second set of values?
set? Explain. (No computing should be necessary.)
16 23 34 56 78 92 93
20 27 38 60 82 96 97

E39. Consider the set of the heights of all female National Collegiate Athletic Association (NCAA) athletes and the set of the heights of all female NCAA basketball players. Which distribution will have the larger mean? Which will have the larger standard deviation? Explain.

E40. Consider the data set 15, 8, 25, 32, 14, 8, 25, and 2. You can replace any one value with a number from 1 to 10. How would you make this replacement
a. to make the standard deviation as large as possible?
b. to make the standard deviation as small as possible?
c. to create an outlier, if possible?

E41. Another measure of center that sometimes is used is the midrange. To find the midrange, compute the mean of the largest value and the smallest value.
The statistics in this computer output summarize the number of viewers of prime-time television shows (in millions) for the week of the last new Seinfeld episode.

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>TrMean</th>
<th>StDev</th>
<th>SEMean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viewers</td>
<td>101</td>
<td>11.187</td>
<td>10.150</td>
<td>9.831</td>
<td>9.896</td>
<td>0.985</td>
</tr>
<tr>
<td>Viewers</td>
<td>Min</td>
<td>Max</td>
<td>Q1</td>
<td>Q3</td>
<td>2.320</td>
<td>76.260</td>
</tr>
</tbody>
</table>

a. Using these summary statistics alone, compute the midrange both with and without the value representing the Seinfeld episode. (Seinfeld had the largest number of viewers and Seinfeld Clips, with 58.53 million viewers, the second largest.) Is the midrange affected much by outliers? Explain.
b. Compute the mean of the ratings without the Seinfeld episode, using only the summary statistics in the computer output.

E42. In computer output like that in E41, TrMean is the trimmed mean. It typically is computed by removing the largest 5% of values and the smallest 5% of values from the data set and then computing the mean of the remaining middle 90% of values. (The percentage that is cut off at each end can vary depending on the software.)
a. Find the trimmed mean of the maximum longevities in Display 2.24 on page 43.
b. Is the trimmed mean affected much by outliers?

E43. This table shows the weights of the pennies in Display 2.3 on page 31. For example, the four pennies in the second interval, 3.0000 g to 3.0199 g, are grouped at the midpoint of this interval, 3.01.

<table>
<thead>
<tr>
<th>Weight</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.99</td>
<td>1</td>
</tr>
<tr>
<td>3.01</td>
<td>4</td>
</tr>
<tr>
<td>3.03</td>
<td>4</td>
</tr>
<tr>
<td>3.05</td>
<td>4</td>
</tr>
<tr>
<td>3.07</td>
<td>7</td>
</tr>
<tr>
<td>3.09</td>
<td>17</td>
</tr>
<tr>
<td>3.11</td>
<td>24</td>
</tr>
<tr>
<td>3.13</td>
<td>17</td>
</tr>
<tr>
<td>3.15</td>
<td>13</td>
</tr>
<tr>
<td>3.17</td>
<td>6</td>
</tr>
<tr>
<td>3.19</td>
<td>2</td>
</tr>
<tr>
<td>3.21</td>
<td>1</td>
</tr>
</tbody>
</table>

a. Find the mean weight of the pennies.
b. Find the standard deviation.
c. Does the standard deviation appear to represent a typical deviation from the mean?

d. Compute the standard deviation using the formula for the standard deviation of values in a frequency table.

E44. Suppose you have five pennies, six nickels, four dimes, and five quarters.
a. Sketch a dot plot of the values of the 20 coins, and use it to estimate the mean.
b. Compute the mean using the formula for the mean of values in a frequency table.
c. Estimate the SD from your plot: Is it closest to 0, 5, 10, 15, or 20?
d. Compute the standard deviation using the formula for the standard deviation of values in a frequency table.
E45. On the first test of the semester, the scores of the first-period class of 30 students had a mean of 75 and a median of 70. The scores of the second-period class of 22 students had a mean of 70 and a median of 68.

a. To the nearest tenth, what is the mean test score of all 52 students? If you cannot calculate the mean of the two classes combined, explain why.

b. What is the median test score of all 52 students? If you cannot find the median of the two classes combined, explain why.

E46. The National Council on Public Polls rebuked the press for its coverage of a Gallup poll of Islamic countries. According to the Council:

News stories based on the Gallup poll reported results in the aggregate without regard to the population of the countries they represent. Kuwait, with less than 2 million Muslims, was treated the same as Indonesia, which has over 200 million Muslims. The “aggregate” quoted in the media was actually the average for the countries surveyed regardless of the size of their populations.

The percentage of people in Kuwait who thought the September 11 terrorist attacks were morally justified was 36%, while the percentage in Indonesia was 4%. [Source: www.ncpp.org.]

a. Suppose that the poll covered only these two countries and that the people surveyed were representative of the entire country. What percentage of all the people in these two countries thought that the terrorist attacks were morally justified?

b. What percentage would have been reported by the press?

2.4 Working with Summary Statistics

Summary statistics are very useful, but only when they are used with good judgment. This section will teach you how to tell which summary statistic to use, how changing units of measurement and the presence of outliers affect your summary statistics, and how to interpret percentiles.

Which Summary Statistic?

Which summary statistics should you use to describe a distribution? Should you use mean and standard deviation? Median and quartiles? Something else? The right choice can depend on the shape of your distribution, so you should always start with a plot. For normal distributions, the mean and standard deviation are nearly always the most suitable. For skewed distributions, the median and quartiles are often the most useful, in part because they have a simple interpretation based on dividing a data set into fourths.

Sometimes, however, the mean and standard deviation will be the right choice even if you have a skewed distribution. For example, if you have a representative sample of house prices for a town and you want to use your sample to estimate the total value of all the town’s houses, the mean is what you want, not the median. In Chapter 7, you’ll see why the mean and standard deviation are the most useful choices when doing statistical inference.
Choosing the right summary statistics is something you will get better at as you build your intuition about the properties of these statistics and how they behave in various situations.

**DISCUSSION**

### Which Summary Statistic?

D21. Explain how to determine the total amount of property taxes if you know the number of houses, the mean value, and the tax rate. In what sense is knowing the mean equivalent to knowing the total?

D22. When a measure of center for the income of a community’s residents is given, that number is usually the median. Why do you think that is the case?

### The Effects of Changing Units

This discussion illustrates some important properties of summary statistics. It will also help you develop your intuition about how the geometry and the arithmetic of working with data are related.

The lowest temperature on record for Washington, D.C., is $-15^\circ F$. How does that temperature compare with the lowest recorded temperatures for capitals of other countries? Display 2.62 gives data for the few capitals whose record low temperatures turn out to be whole numbers on both the Fahrenheit and Celsius scales.

<table>
<thead>
<tr>
<th>City</th>
<th>Country</th>
<th>Temperature ($^\circ F$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addis Ababa</td>
<td>Ethiopia</td>
<td>32</td>
</tr>
<tr>
<td>Algiers</td>
<td>Algeria</td>
<td>32</td>
</tr>
<tr>
<td>Bangkok</td>
<td>Thailand</td>
<td>50</td>
</tr>
<tr>
<td>Madrid</td>
<td>Spain</td>
<td>14</td>
</tr>
<tr>
<td>Nairobi</td>
<td>Kenya</td>
<td>41</td>
</tr>
<tr>
<td>Brazilia</td>
<td>Brazil</td>
<td>32</td>
</tr>
<tr>
<td>Warsaw</td>
<td>Poland</td>
<td>$-22$</td>
</tr>
</tbody>
</table>

Display 2.62 Record low temperatures for seven capitals.

[Source: National Climatic Data Center, 2002.]

The dot plot in Display 2.63 shows that the temperatures are centered at about $32^\circ F$, with an outlier at $-22^\circ F$. The spread and shape are hard to determine with only seven values.

Display 2.63 Dot plot for record low temperatures in degrees Fahrenheit for seven capitals.
What happens to the shape, center, and spread of this distribution if you convert each temperature to the number of degrees above or below freezing, 32°F? To find out, subtract 32 from each value, and plot the new values. Display 2.64 shows that the center of the dot plot is now at 0 rather than 32 but the spread and shape are unchanged.

Adding (or subtracting) a constant to each value in a set of data doesn’t change the spread or the shape of a distribution but slides the entire distribution a distance equivalent to the constant. Thus, the transformation amounts to a recentering of the distribution.

What happens to the shape and spread of this distribution if you convert each temperature to degrees Celsius? The Celsius scale measures temperature based on the number of degrees above or below freezing, but it takes 1.8°F to make 1°C. To convert, divide each value in Display 2.64 by 1.8 (or, equivalently, multiply by \( \frac{1}{1.8} \)), and plot the new values. Display 2.65 shows that the center of the new dot plot is still at 0 and the shape is the same. However, the spread has shrunk by a factor of \( \frac{1}{1.8} \).

Multiplying each value in a set of data by a positive constant doesn’t change the basic shape of the distribution. Both the mean and the spread are multiplied by that number. This transformation amounts to a rescaling of the distribution.

[See Calculator Note 2G to explore on your calculator the effects of changing units.]

### Recentering and Rescaling a Data Set

**Recentering** a data set—adding the same number \( c \) to all the values in the set—doesn’t change the shape or spread but slides the entire distribution by the amount \( c \), adding \( c \) to the median and the mean.

**Rescaling** a data set—multiplying all the values in the set by the same positive number \( d \)—doesn’t change the basic shape but stretches or shrinks the distribution, multiplying the spread (IQR or standard deviation) by \( d \) and multiplying the center (median or mean) by \( d \).
The Effects of Changing Units


a. A set of prices, in U.S. dollars, has mean $20 and standard deviation $5. Find the mean and standard deviation of the prices expressed in pesos.

b. Another set of prices, in Mexican pesos, has a median of 94 pesos and quartiles of 47 pesos and 188 pesos. Find the median and quartiles of the same prices expressed in U.S. dollars.

The Influence of Outliers

A summary statistic is resistant to outliers if the summary statistic is not changed very much when an outlier is removed from the set of data. If the summary statistic tends to be affected by the removal of outliers, it is sensitive to outliers.

Display 2.66 shows a dot plot of the number of viewers of prime-time television shows (in millions) in a particular week. (A boxplot of these data is shown in Display 2.55 on page 70.) The three highest values—the three shows with the largest numbers of viewers—are outliers.

The printout in Display 2.67 gives the summary statistics for all 101 shows.

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>StDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratings</td>
<td>101</td>
<td>11.187</td>
<td>10.150</td>
<td>9.896</td>
</tr>
</tbody>
</table>

The printout in Display 2.68 gives the summary statistics for the number of viewers when the three outliers are removed from the set of data. Compare this printout with the one in Display 2.67 and notice which summary statistics are most sensitive to the outliers.

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>StDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Outs</td>
<td>98</td>
<td>9.666</td>
<td>10.145</td>
<td>4.250</td>
</tr>
</tbody>
</table>

Display 2.67 Printout of summary statistics for number of viewers.

Display 2.68 Summary statistics for number of viewers without outliers.
### The Influence of Outliers

D24. Are these measures of center for the number of television viewers affected much by the three outliers? (Refer to Displays 2.66–2.68.) Explain.

a. mean  

b. median  

d25. Are these measures of spread for the number of television viewers affected much by the three outliers? Explain why or why not.

a. range  

b. standard deviation  

c. interquartile range

### Percentiles and Cumulative Relative Frequency Plots

Percentiles measure position within a data set. The first quartile, $Q_1$, of a distribution is the 25th percentile—the value that separates the lowest 25% of the ordered values from the rest. The median is the 50th percentile, and $Q_3$ is the 75th percentile. You can define other percentiles in the same way. The 10th percentile, for example, is the value that separates the lowest 10% of ordered values in a distribution from the rest. In general, a value is at the $k$th percentile if $k\%$ of all values are less than or equal to it.

For large data sets, you might see data listed in a table or plotted in a graph, like those for the SAT I critical reading scores in Display 2.69. Such a plot is sometimes called a **cumulative percentage plot** or a **cumulative relative frequency plot**. The table shows that, for example, 28% of the students received a score of 450 or lower and about 13% received a score between 400 and 450.

<table>
<thead>
<tr>
<th>Score</th>
<th>Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>800</td>
<td>99+</td>
</tr>
<tr>
<td>750</td>
<td>96</td>
</tr>
<tr>
<td>700</td>
<td>93</td>
</tr>
<tr>
<td>650</td>
<td>87</td>
</tr>
<tr>
<td>600</td>
<td>76</td>
</tr>
<tr>
<td>550</td>
<td>62</td>
</tr>
<tr>
<td>500</td>
<td>45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Score</th>
<th>Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>450</td>
<td>28</td>
</tr>
<tr>
<td>400</td>
<td>15</td>
</tr>
<tr>
<td>350</td>
<td>7</td>
</tr>
<tr>
<td>300</td>
<td>3</td>
</tr>
<tr>
<td>250</td>
<td>1</td>
</tr>
<tr>
<td>200</td>
<td>—</td>
</tr>
</tbody>
</table>

[See Calculator Note 2H to learn how to construct cumulative relative frequency plots on your calculator.]

### DISCUSSION

#### Percentiles and Cumulative Relative Frequency Plots

**D26.** Refer to Display 2.69.

- Use the plot to estimate the percentile for an SAT I critical reading score of 425.
- What two values enclose the middle 90% of SAT scores? The middle 95%?
- Use the table to estimate the score that falls at the 40th percentile.

**D27.** What proportion of cases lie between the 5th and 95th percentiles of a distribution? What percentiles enclose the middle 95% of the cases in a distribution?

#### Summary 2.4: Working with Summary Statistics

Knowing which summary statistic to use depends on what use you have for that summary statistic.

- If a summary statistic doesn’t change much whether you include or exclude outliers from your data set, it is said to be resistant to outliers.
  - The median and quartiles are resistant to outliers.
  - The mean and standard deviation are sensitive to outliers.

Recentering a data set—adding the same number \(c\) to all the values—slides the entire distribution. It doesn’t change the shape or spread but adds \(c\) to the median and the mean. Rescaling a data set—multiplying all the values by the same nonzero number \(d\)—is like stretching or squeezing the distribution. It doesn’t change the basic shape but multiplies the spread (IQR or standard deviation) by \(|d|\) and multiplies the measure of center (median or mean) by \(d\).

The percentile of a value tells you what percentage of all values lie at or below the given value. The 30th percentile, for example, is the value that separates the distribution into the lowest 30% of values and the highest 70% of values.

#### Practice

**Which Summary Statistic?**

**P26.** A community in Nevada has 9751 households, with a median house price of $320,000 and a mean price of $392,059.

- Why is the mean larger than the median?
- The property tax rate is about 1.15%.
  - What total amount of taxes will be assessed on these houses?
- What is the average amount of taxes per house?

**P27.** A news release at www.polk.com stated that the median age of cars being driven in 2004 was 8.9 years, the oldest to date. The median was 8.3 years in 2000 and 7.7 years in 1995.

- Why were medians used in this news story?
- What reasons might there be for the increase in the median age of cars? (The median age in 1970 was only 4.9 years!)
The Effects of Changing Units

P28. The mean height of a class of 15 children is 48 in., the median is 45 in., the standard deviation is 2.4 in., and the interquartile range is 3 in. Find the mean, standard deviation, median, and interquartile range if
   a. you convert each height to feet
   b. each child grows 2 in.
   c. each child grows 4 in. and you convert the heights to feet

P29. Compute the means and standard deviations (use the formula for $s$) of these sets of numbers. Use recentering and rescaling wherever you can to avoid or simplify the arithmetic.
   a. 1 2 3
   b. 11 12 13
   c. 10 20 30
   d. 105 110 115
   e. $-800 \quad -900 \quad -1000$

The Influence of Outliers

P30. The histogram and boxplot in Display 2.70 and the summary statistics in Display 2.71 show the record low temperatures for the 50 states.
   a. Hawaii has a lowest recorded temperature of 12°F. The boxplot shows Hawaii as an outlier. Verify that this is justified.
   b. Suppose you exclude Hawaii from the data set. Copy the table in Display 2.71, substituting the value (or your best estimate if you don’t have enough information to compute the value) of each summary statistic with Hawaii excluded.


<table>
<thead>
<tr>
<th>Summary of Lowest Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Median</td>
</tr>
<tr>
<td>StdDev</td>
</tr>
<tr>
<td>Min</td>
</tr>
<tr>
<td>Max</td>
</tr>
<tr>
<td>Range</td>
</tr>
<tr>
<td>Lower 25 %tile</td>
</tr>
<tr>
<td>Upper 75 %tile</td>
</tr>
</tbody>
</table>

Display 2.71 Summary statistics for lowest temperatures for the 50 states.

Percentiles and Cumulative Relative Frequency Plots

P31. Estimate the quartiles and the median of the SAT I critical reading scores in Display 2.69 on page 78, and then use these values to draw a boxplot of the distribution. What is the IQR?

Exercises

E47. Discuss whether you would use the mean or the median to measure the center of each set of data and why you prefer the one you chose.
   a. the prices of single-family homes in your neighborhood
   b. the yield of corn (bushels per acre) for a sample of farms in Iowa
   c. the survival time, following diagnosis, of a sample of cancer patients
E48. **Mean versus median.**

a. You are tracing your family tree and would like to go back to the year 1700. To estimate how many generations back you will have to trace, would you need to know the median length of a generation or the mean length of a generation?

b. If a car trip takes 3 h, do you need to know the mean speed or the median speed in order to find the total distance?

c. Suppose all trees in a forest are right circular cylinders with radius 3 ft. The heights vary, but the mean height is 45 ft, the median is 43 ft, the IQR is 3 ft, and the standard deviation is 3.5 ft. From this information, can you compute the total volume of wood in all the trees?

E49. The histogram in Display 2.72 shows record high temperatures for the 50 states.

![Histogram](image)


a. Suppose each temperature is converted from degrees Fahrenheit, \( T \), to degrees Celsius, \( C \), using the formula

\[
C = \frac{5}{9}(T - 32)
\]

If you make a histogram of the temperatures in degrees Celsius, how will it differ from the one in Display 2.72?

b. The summary statistics in Display 2.73 are for record high temperatures in degrees Fahrenheit. Make a similar table for the temperatures in degrees Celsius.

```latex
\begin{tabular}{|l|l|l|l|l|}
\hline
\textbf{Variable} & \textbf{N} & \textbf{Mean} & \textbf{Median} & \textbf{StDev} \\
\hline
HighTemp & 50 & 114.10 & 114.00 & 6.69 \\
\hline
\end{tabular}
```

**Display 2.73** Summary statistics for record high temperatures for the 50 U.S. states.

c. Are there any outliers in the data in °C?

E50. Tell how you could use recentering and rescaling to simplify the computation of the mean and standard deviation for this list of numbers:

\[5478.1 \ 5478.3 \ 5478.3 \ 5478.9 \ 5478.4 \ 5478.2\]

E51. Suppose a constant \( c \) is added to each value in a set of data, \( x_1, x_2, x_3, x_4, \) and \( x_5 \). Prove that the mean increases by \( c \) by comparing the formula for the mean of the original data to the formula for the mean of the recentered data.

E52. Suppose a constant \( c \) is added to each value in a set of data, \( x_1, x_2, x_3, x_4, \) and \( x_5 \). Prove that the standard deviation is unchanged by comparing the formula for the standard deviation of the original data to the formula for the standard deviation of the recentered data.

E53. The cumulative relative frequency plot in Display 2.74 shows the amount of change carried by a group of 200 students. For example, about 80% of the students had $0.75 or less.

![Cumulative Percentage Plot](image)

**Display 2.74** Cumulative percentage plot of amount of change.

a. From this plot, estimate the median amount of change.
b. Estimate the quartiles and the interquartile range.

c. Is the original set of amounts of change skewed right, skewed left, or symmetric?

d. Does the data set look as if it should be modeled by a normal distribution? Explain your reasoning.

E54. Use Display 2.74 to make a boxplot of the amounts of change carried by the students.

E55. Did you ever wonder how speed limits on roadways are determined? Most government jurisdictions set speed limits by this standard practice, described on the website of the Michigan State Police.

Speed studies are taken during times that represent normal free-flow traffic. Since modified speed limits are the maximum allowable speeds, roadway conditions must be close to ideal. The primary basis for establishing a proper, realistic speed limit is the nationally recognized method of using the 85th percentile speed. This is the speed at or below which 85% of the traffic moves. [Source: www.michigan.gov.]

The 85th percentile speed typically is rounded down to the nearest 5 miles per hour. The table and histogram in Display 2.75 give the measurements of the speeds of 1000 cars on a stretch of road in Mellowville with no curviness or other additional factors. At what speed would the speed limit be set if the guidelines described were followed?

<table>
<thead>
<tr>
<th>Speed (mi/h)</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>2</td>
</tr>
<tr>
<td>26</td>
<td>31</td>
</tr>
<tr>
<td>27</td>
<td>92</td>
</tr>
<tr>
<td>28</td>
<td>149</td>
</tr>
<tr>
<td>29</td>
<td>178</td>
</tr>
<tr>
<td>30</td>
<td>156</td>
</tr>
<tr>
<td>31</td>
<td>157</td>
</tr>
<tr>
<td>32</td>
<td>99</td>
</tr>
<tr>
<td>33</td>
<td>74</td>
</tr>
<tr>
<td>34</td>
<td>31</td>
</tr>
<tr>
<td>35</td>
<td>16</td>
</tr>
<tr>
<td>36</td>
<td>13</td>
</tr>
<tr>
<td>37</td>
<td>1</td>
</tr>
<tr>
<td>38</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1000</strong></td>
</tr>
</tbody>
</table>

Display 2.75  Speed of 1000 cars in Mellowville.

E56. Refer to the distribution of speeds in E55. Make a cumulative relative frequency plot of these speeds.

E57. The cumulative relative frequency plot in Display 2.76 gives the ages of the CEOs (Chief Executive Officers) of the 500 largest U.S. companies. Does A, B, or C give its median and quartiles? Using the diagram, explain why your choice is correct.

A. $Q_1$, 51; median 56; $Q_3$, 60
B. $Q_1$, 50; median 60; $Q_3$, 70
C. $Q_1$, 25; median 50; $Q_3$, 75
2.5 The Normal Distribution

You have seen several reasons why the normal distribution is so important:

- It tells you how variability in repeated measurements often behaves (diameters of tennis balls).
- It tells you how variability in populations often behaves (weights of pennies, SAT scores).
- It tells you how means (and some other summary statistics) computed from random samples behave (the Westvaco case, Activity 1.2a).

In this section, you will learn that if you know that a distribution is normal (shape), then the mean (center) and standard deviation (spread) tell you everything else about the distribution. The reason is that, whereas skewed distributions come in many different shapes, there is only one normal shape. It's true that one normal distribution might appear tall and thin while another looks short and fat. However, the x-axis of the tall, thin distribution can be stretched out so that it looks exactly the same as the short, fat one.

The Standard Normal Distribution

Because all normal distributions have the same basic shape, you can use recentering and rescaling to change any normal distribution to the one with mean 0 and standard deviation 1. Solving problems involving normal distributions depends on this important property.

The normal distribution with mean 0 and standard deviation 1 is called the standard normal distribution. In this distribution, the variable along the horizontal axis is called a z-score.

Display 2.76  Cumulative relative frequency plot of CEO ages. [Source: www.forbes.com]
The standard normal distribution is symmetric, with total area under the curve equal to 1, or 100%. To find the percentage, \( P \), that describes the area to the left of the corresponding \( z \)-score, you can use the \( z \)-table or your calculator.

The next two examples show you how to use the \( z \)-table, Table A on pages 824–825.

**Example: Finding the Percentage When You Know the \( z \)-Score**

Find the percentage, \( P \), of values less than \( z = 1.23 \), the shaded area in Display 2.77. Find the percentage greater than \( z = 1.23 \).

**Solution**

Think of 1.23 as 1.2 + 0.03. In Table A on pages 824–825, find the row labeled 1.2 and the column headed .03. Where this row and column intersect, you find the number .8907. That means that 89.07% of standard normal scores are less than 1.23.

The total area under the curve is 1, so the proportion of values greater than \( z = 1.23 \) is \( 1 - 0.8907 \), or 0.1093, which is 10.93%.

A graphing calculator will give you greater accuracy in finding the proportion of values that lie between two specified values in a standard normal distribution. For example, you can find the proportion of values that are less than 1.23 in a standard normal distribution like this:

![Graphing calculator output](image)

[To learn more about calculating the proportion of values between two \( z \)-scores, see Calculator Note 21.]

**Example: Finding the \( z \)-Score When You Know the Percentage**

Find the \( z \)-score that falls at the 75th percentile of the standard normal distribution, that is, the \( z \)-score that divides the bottom 75% of values from the rest.
**Solution**

First make a sketch of the situation, as in Display 2.78.

![Display 2.78](image)

**Display 2.78** The z-score that corresponds to the 75th percentile.

Look for .7500 in the body of Table A. No value in the table is exactly equal to .7500. The closest value is .7486. The value .7486 sits at the intersection of the row labeled .60 and the column headed .07, so the corresponding z-score is roughly 0.60 + 0.07, or 0.67.

You can use a graphing calculator to find the 75th percentile of a standard normal distribution like this:

![Graphing Calculator](image)

[To learn more about finding the z-score that has a specified proportion of values below it, see Calculator Note 2].

---

**DISCUSSION**

**The Standard Normal Distribution**

D28. For the standard normal distribution,

a. what is the median?

b. what is the lower quartile?

c. what z-score falls at the 95th percentile?

d. what is the IQR?

**Standard Units: How Many Standard Deviations Is It from Here to the Mean?**

Converting to standard units, or standardizing, is the two-step process of recentering and rescaling that turns any normal distribution into the standard normal distribution.

First you recenter all the values of the normal distribution by subtracting the mean from each. This gives you a distribution with mean 0. Then you rescale by...
dividing all the values by the standard deviation. This gives you a distribution with standard deviation 1. You now have a standard normal distribution. You can also think of the two-step process of standardizing as answering two questions: How far above or below the mean is my score? How many standard deviations is that?

The **standard units** or **z-score** is the number of standard deviations that a given $x$-value lies above or below the mean.

How far and which way to the mean?  
\[ x - \text{mean} \]

How many standard deviations is that?  
\[ z = \frac{x - \text{mean}}{SD} \]

---

**Example: Computing a z-Score**

In a recent year, the distribution of SAT I math scores for the incoming class at the University of Georgia was roughly normal, with mean 610 and standard deviation 69. What is the z-score for a University of Georgia student who got 560 on the math SAT?

**Solution**

A score of 560 is 50 points below the mean of 610. This is \( \frac{-50}{69} \) or 0.725 standard deviation below the mean. Alternatively, using the formula,

\[ z = \frac{x - \text{mean}}{SD} = \frac{560 - 610}{69} \approx -0.725 \]

So the student’s z-score is \(-0.725\).

---

To “unstandardize,” think in reverse. Alternatively, you can solve the z-score formula for $x$ and get

\[ x = \text{mean} + z \cdot SD \]

**Example: Finding the Value When You Know the z-Score**

What was a University of Georgia student’s SAT I math score if his or her score was 1.6 standard deviations above the mean?

**Solution**

The score that is 1.6 standard deviations above the mean is

\[ x = \text{mean} + z \cdot SD = 610 + 1.6(69) = 720 \]
Example: Using z-Scores to Make a Comparison

In the United States, heart disease kills roughly one-and-a-quarter times as many people as cancer. If you look at the death rate per 100,000 residents by state, the distributions for the two diseases are roughly normal, provided you leave out Alaska and Utah, which are outliers because of their unusually young populations. The means and standard deviations for all 50 states are given here.

<table>
<thead>
<tr>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart disease</td>
<td>238</td>
</tr>
<tr>
<td>Cancer</td>
<td>196</td>
</tr>
</tbody>
</table>

Alaska had 88 deaths per 100,000 residents from heart disease, and 111 from cancer. Explain which death rate is more extreme compared to other states. [Source: Centers for Disease Control, National Vital Statistics Report, vol. 53, no. 5, October 12, 2004.]

Solution

\[ z_{\text{heart}} = \frac{88 - 238}{52} \approx -2.88 \]

\[ z_{\text{cancer}} = \frac{111 - 196}{31} \approx -2.74 \]

Alaska’s death rate for heart disease is 2.88 standard deviations below the mean. The death rate for cancer is 2.74 standard deviations below the mean. These rates are about equally extreme, but the death rate for heart disease is slightly more extreme.

DISCUSSION

Standard Units

D29. Standardizing is a process that is similar to other processes you have seen already.

a. You’re driving at 60 mi/h on the interstate and are now passing the marker for mile 200, and your exit is at mile 80. How many hours from your exit are you?

b. What two arithmetic operations did you do to get the answer in part a? Which operation corresponds to recentering? Which corresponds to rescaling?

Solving the Unknown Percentage Problem and the Unknown Value Problem

Now you know all you need to know to analyze situations involving two related problems concerning a normal distribution: finding a percentage when you know the value, and finding the value when you know the percentage.
Example: Percentage of Males Taller Than 74 Inches

For groups of similar individuals, heights often are approximately normal in their distribution. For example, the heights of 18- to 24-year-old males in the United States are approximately normal, with mean 70.1 in. and standard deviation 2.7 in. What percentage of these males are more than 74 in. tall? [Source: U.S. Census Bureau, *Statistical Abstract of the United States*, 1991.]

Solution

First make a sketch of the situation, as in Display 2.79. Draw a normal shape above a horizontal axis. Place the mean in the middle on the axis. Then mark and label the points that are two standard deviations either side of the mean, 64.7 and 75.5, so that about 95% of the values lie between them. Next, mark and label the points that are one and three standard deviations either side of the mean (67.4 and 72.8, and 62 and 78.2). Finally, estimate the location of the given value of $x$ and mark it on the axis.

Display 2.79  The percentage of heights greater than 74 in.

Standardize:

\[ z = \frac{x - \text{mean}}{\text{SD}} = \frac{74 - 70.1}{2.7} = 1.44 \]

Look up the proportion: The area to the left of the $z$-score 1.44 is 0.9251, so the proportion of males taller than 74 in. is $1 - 0.9251$, or 0.0749 or 7.49%.

Example: Percentage of Males Between 72 and 74 Inches Tall

The heights of 18- to 24-year-old males in the United States are approximately normal, with mean 70.1 in. and standard deviation 2.7 in. What percentage of these males are between 72 and 74 in. tall?
Solution

First make a sketch, as in Display 2.80.

Display 2.80 The percentage of male heights between 72 and 74 in.

Standardize: From the previous example, a height of 74 in. has a z-score of 1.44. For a height of 72 in.,

\[ z = \frac{x - \text{mean}}{\text{SD}} = \frac{72 - 70.1}{2.7} \approx 0.70 \]

Look up the proportion: The area to the left of the z-score 1.44 is 0.9251. The area to the left of the z-score 0.70 is 0.7580. The area you want is the area between these two z-scores, which is 0.9251 - 0.7580, or 0.1671. So the percentage of 18- to 24-year-old males between 72 and 74 in. tall is about 16.71%.

You can also use a graphing calculator to find this value. [See Calculator Note 2I for more details.]

Example: 75th Percentile of Female Heights

The heights of females in the United States who are between the ages of 18 and 24 are approximately normally distributed, with mean 64.8 in. and standard deviation 2.5 in. What height separates the shortest 75% from the tallest 25%?
Solution

First make a sketch, as in Display 2.81.

\[
\begin{array}{c}
\text{Heights (in.)} \\
59.8 \quad 64.8 \quad 69.8
\end{array}
\]

\[
P = 0.75
\]

Display 2.81 The 75th percentile in height for women age 18 to 24.

Look up the \( z \)-score: If the proportion \( P \) is 0.75, then from Table A, you find that \( z \) is approximately 0.67.

Unstandardize:

\[
x = \text{mean} + z \cdot SD = 64.8 + 0.67(2.5) = 66.475 \text{ in.}
\]

For an unknown percentage problem:

First standardize by converting the given value to a \( z \)-score:

\[
z = \frac{x - \text{mean}}{SD}
\]

Then look up the percentage.

For an unknown value problem, reverse the process:

First look up the \( z \)-score corresponding to the given percentage. Then unstandardize:

\[
x = \text{mean} + z \cdot SD
\]

DISCUSSION

Solving the Unknown Percentage Problem and the Unknown Value Problem

D30. Age of cars. The cars in Clunkerville have a mean age of 12 years and a standard deviation of 8 years. What percentage of cars are more than 4 years old? (Warning: This is a trick question.)

Central Intervals for Normal Distributions

You learned in Section 2.1 that if a distribution is roughly normal, about 68% of the values lie within one standard deviation of the mean. It is helpful to memorize this fact as well as the others in the box on the next page.
Central Intervals for Normal Distributions

68% of the values lie within 1 standard deviation of the mean.

90% of the values lie within 1.645 standard deviations of the mean.

95% of the values lie within 1.96 (or about 2) standard deviations of the mean.

99.7% (or almost all) of the values lie within 3 standard deviations of the mean.

Example: Middle 90% of Death Rates from Cancer

According to the table on page 87, the death rates per 100,000 residents from cancer are approximately normal, with mean 196 and SD 31. The middle 90% of death rates are between what two numbers?

Solution

The middle 90% of values in this distribution lie within 1.645 standard deviations of the mean, 196. That is, about 90% of the values lie in the interval $196 \pm 1.645(31)$, or between about 145 and 247.
You can confirm this result using your calculator: Shade the area under this normal curve between 145 and 247 and calculate the area. [See Calculator Note 2K.]

Central Intervals for Normal Distributions

D31. Use Table A on pages 824–825 to verify that 99.7% of the values in a normal distribution lie within three standard deviations of the mean.

Summary 2.5: The Normal Distribution

The standard normal distribution has mean 0 and standard deviation 1. All normal distributions can be converted to the standard normal distribution by converting to standard units:

- First, recenter by subtracting the mean.
- Then rescale by dividing by the standard deviation:

\[ z = \frac{x - \text{mean}}{SD} \]

Standard units \( z \) tell how far a value \( x \) is from the mean, measured in standard deviations. If you know \( z \), you can find \( x \) by using the formula \( x = \text{mean} + z \cdot SD \).

If your population is approximately normal, you can compute \( z \) and then use Table A or your calculator to find the corresponding proportion. Be sure to make a sketch so that you know whether to use the proportion in the table or to subtract that proportion from 1.

For any normal distribution,

- 68% of the values lie within 1 standard deviation of the mean
- 90% of the values lie within 1.645 standard deviations of the mean
- 95% of the values lie within 1.960 (or about 2) standard deviations of the mean
- 99.7% (or almost all) of the values lie within 3 standard deviations of the mean

DISCUSSION

Practice

The Standard Normal Distribution

P32. Find the percentage of values below each given \( z \)-score in a standard normal distribution.

a. \(-2.23\)  b. \(-1.67\)  c. \(-0.40\)  d. \(0.80\)

P33. Find the \( z \)-score that has the given percentage of values below it in a standard normal distribution.

a. 32%  b. 41%  c. 87%  d. 94%
P34. What percentage of values in a standard normal distribution fall between
   a. −1.46 and 1.46?
   b. −3 and 3?

P35. For a standard normal distribution, what interval contains
   a. the middle 90% of z-scores?
   b. the middle 95% of z-scores?

Standard Units
P36. Refer to the table in the example on page 87.
   a. California had 196 deaths from heart disease and 154 deaths from cancer per 100,000 residents. Which rate is more extreme compared to other states? Why?
   b. Florida had 295 deaths from heart disease and 234 deaths from cancer per 100,000 residents. Which rate is more extreme?
   c. Colorado had an unusually low rate of heart disease, 143 deaths per 100,000 residents. Hawaii had an unusually low rate of cancer, 156 deaths per 100,000 residents. Which is more extreme?

Solving the Unknown Percentage Problem and the Unknown Value Problem
P37. The heights of 18- to 24-year-old males in the United States are approximately normal, with mean 70.1 in. and standard deviation 2.7 in. The heights of 18- to 24-year-old females are also approximately normally distributed and have mean 64.8 in. and standard deviation 2.5 in.
   a. Estimate the percentage of U.S. males between 18 and 24 who are 6 ft tall or taller.
   b. How tall does a U.S. woman between 18 and 24 have to be in order to be at the 35th percentile of heights?

Central Intervals for Normal Distributions
P38. Refer to the table in the example on page 87.
   a. The middle 90% of the states’ death rates from heart disease fall between what two numbers?
   b. The middle 68% of death rates from heart disease fall between what two numbers?

P39. Refer to the information in P37. Which of the following heights are outside the middle 95% of the distribution? Which are outside the middle 99%?
   A. a male who is 79 in. tall
   B. a female who is 68 in. tall
   C. a male who is 65 in. tall
   D. a female who is 65 in. tall

Exercises
E59. What percentage of values in a standard normal distribution fall
   a. below a z-score of 1.00? 2.53?
   b. below a z-score of −1.00? −2.53?
   c. above a z-score of −1.5?
   d. between z-scores of −1 and 1?

E60. On the same set of axes, draw two normal curves with mean 50, one having standard deviation 5 and the other having standard deviation 10.

E61. Standardizing. Convert each of these values to standard units, z. (Do not use a calculator. These are meant to be done in your head.)
   a. x = 12, mean 10, SD 1
   b. x = 12, mean 10, SD 2
   c. x = 12, mean 9, SD 2
   d. x = 12, mean 9, SD 1
   e. x = 7, mean 10, SD 3
   f. x = 5, mean 10, SD 2

E62. Unstandardizing. Find the value of x that was converted to the given z-score.
   a. z = 2, mean 20, SD 5
   b. z = −1, mean 25, SD 3
   c. z = −1.5, mean 100, SD 10
   d. z = 2.5, mean −10, SD 0.2
E63. SAT I critical reading scores are scaled so that they are approximately normal, with mean about 505 and standard deviation about 111.
   a. Find the probability that a randomly selected student has an SAT I critical reading score
      i. between 400 and 600
      ii. over 700
      iii. below 450
   b. What SAT I critical reading scores fall in the middle 95% of the distribution?

E64. SAT I math scores are scaled so that they are approximately normal, with mean about 511 and standard deviation about 112. A college wants to send letters to students scoring in the top 20% on the exam. What SAT I math score should the college use as the dividing line between those who get letters and those who do not?

E65. Height limitations for flight attendants.
   To work as a flight attendant for United Airlines, you must be between 5 ft 2 in. and 6 ft tall. [Source: www.ual.com.] The mean height of 18- to 24-year-old males in the United States is about 70.1 in., with a standard deviation of 2.7 in. The mean height of 18- to 24-year-old females is about 64.8 in., with a standard deviation of 2.5 in. Both distributions are approximately normal. What percentage of men this age meet the height limitation? What percentage of women this age meet the height limitation?

E66. Where is the next generation of male professional basketball players coming from?
   a. The mean height of 18- to 24-year-old males in the United States is approximately normally distributed, with mean 70.1 in. and standard deviation 2.7 in. Use this information to approximate the percentage of men in the United States between the ages of 18 and 24 who are as tall as or taller than each basketball player listed here. Then, using the fact that there are about 13 million men between the ages of 18 and 24 in the United States, estimate how many are as tall as or taller than each player.
      i. Shawn Marion, 6 ft 7 in.
      ii. Allen Iverson, 6 ft 0 in.
      iii. Shaquille O’Neal, 7 ft 1 in.
   b. Distributions of real data that are approximately normal tend to have heavier “tails” than the ideal normal curve. Does this mean your estimates in part a are too small, too big, or just right?

E67. Puzzle problems. Problems that involve computations with the normal distribution have four quantities: mean, standard deviation, value \( x \), and proportion \( P \) below value \( x \). Any three of these values are enough to determine the fourth. Think of each row in this table as little puzzles, and find the missing value in each case. This isn’t the sort of thing you are likely to run into in practice, but solving the puzzles can help you become more skilled at working with the normal distribution.

<table>
<thead>
<tr>
<th>Mean</th>
<th>SD</th>
<th>( x )</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>a</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>b</td>
<td>0.18</td>
</tr>
<tr>
<td>c</td>
<td>3</td>
<td>6</td>
<td>0.09</td>
</tr>
<tr>
<td>10</td>
<td>d</td>
<td>12</td>
<td>0.60</td>
</tr>
</tbody>
</table>

E68. More puzzle problems. In each row of this table, assume the distribution is normal. Knowing any two of the mean, standard deviation, \( Q_1 \), and \( Q_3 \) is enough to determine the other two. Complete the table.

<table>
<thead>
<tr>
<th>Mean</th>
<th>SD</th>
<th>( Q_1 )</th>
<th>( Q_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>5</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>c</td>
<td>d</td>
<td>120</td>
<td>180</td>
</tr>
<tr>
<td>e</td>
<td>10</td>
<td>100</td>
<td>f</td>
</tr>
<tr>
<td>10</td>
<td>g</td>
<td>h</td>
<td>11</td>
</tr>
</tbody>
</table>

E69. ACT scores are approximately normally distributed, with mean 18 and standard deviation 6. Without using your calculator, roughly what percentage of scores are between 12 and 24? Between 6 and 30? Above 24? Below 24? Above 6? Below 6?

E70. A group of subjects tested a certain brand of foam earplug. The number of decibels (dB) that noise was reduced for these subjects was
approximately normally distributed, with mean 30 dB and standard deviation 3.6 dB. The middle 95% of noise reductions were between what two values?

E71. The heights of 18- to 24-year-old males in the United States are approximately normally distributed with mean 70.1 in. and standard deviation 2.7 in.

a. If you select a U.S. male between ages 18 and 24 at random, what is the approximate probability that he is less than 68 in. tall?

b. There are roughly 13 million 18- to 24-year-old males in the United States. About how many are between 67 and 68 in. tall?

c. Find the height of 18- to 24-year-old males that falls at the 90th percentile.

E72. If the measurements of height are transformed from inches into feet, will that change the shape of the distribution in E71? Describe the distribution of male heights in terms of feet rather than inches.

E73. The British monarchy. Over the 1200 years of the British monarchy, the average reign of kings and queens has lasted 18.5 years, with a standard deviation of 15.4 years.

a. What can you say about the shape of the distribution based on the information given?

b. Suppose you made the mistake of assuming a normal distribution. What fraction of the reigns would you estimate lasted a negative number of years?

c. Use your work in part b to suggest a rough rule for using the mean and standard deviation of a set of positive values to check whether it is possible that a distribution might be approximately normal.

E74. NCAA scores. The histogram in Display 2.82 was constructed from the total of the scores of both teams in all NCAA basketball play-off games over a 57-year period.

a. Approximate the mean of this distribution.

b. Approximate the standard deviation of this distribution.

c. Between what two values do the middle 95% of total points scored lie?

d. Suppose you choose a game at random from next year’s NCAA play-offs. What is the approximate probability that the total points scored in this game will exceed 150? 190? Do you see any potential weaknesses in your approximations?

Chapter Summary

Distributions come in various shapes, and the appropriate summary statistics (for center and spread) usually depend on the shape, so you should always start with a plot of your data.

Common symmetric shapes include the uniform (rectangular) distribution and the normal distribution. There are also various skewed distributions. Bimodal distributions often result from mixing cases of two kinds.
Dot plots, stemplots, and histograms show distributions graphically and let you estimate center and spread visually from the plot.

For approximately normal distributions, you ordinarily use the mean (balance point) and standard deviation as the measure of center and spread. If you know the mean and standard deviation of a normal distribution, you can use z-scores and Table A or your calculator to find the percentage of values in any interval.

The mean and standard deviation are not resistant—their values are sensitive to outliers. For a description of a skewed distribution, you should consider using the median (halfway point) and quartiles (medians of the lower and upper halves of the data) as summary statistics.

Later on, when you make inferences about the entire population from a sample taken from that population, the sample mean and standard deviation will be the most useful summary statistics, even if the population is skewed.

**Review Exercises**

E75. The map in Display 2.83, from the U.S. National Weather Service, gives the number of tornadoes by state, including the District of Columbia.

![Display 2.83](Image)

Display 2.83 The number of tornadoes per state in a recent year. [Source: www.ncdc.noaa.gov.]

a. Make a stemplot of the number of tornadoes.
b. Write the five-number summary.
c. Identify any outliers.
d. Draw a boxplot.
e. Compare the information in your stemplot with the information in your boxplot. Which plot is more informative?
f. Describe the shape, center, and spread of the distribution of the number of tornadoes.

E76. Display 2.84 shows some results of the Third International Mathematics and Science study for various countries. Each case is a school.

![Display 2.84](Image)

Display 2.84 Boxplots of mathematics instruction time by country for 9-year-olds. [Source: Report #8, April 1998, of the Third International Mathematics and Science Study (TIMSS), p. 6.]

a. Estimate the median for the United States. Use this median value in a sentence that makes it clear what the median represents in this context.
b. Why are there only lines and no boxes for Norway and Singapore?
c. Describe how the distribution of values for the United States compares to the distributions of values for the other countries.

E77. A university reports that the middle 50% of the SAT I math scores of its students were between 585 and 670, with half the scores above 605 and half below.
   a. What SAT I math scores would be considered outliers for that university?
   b. What can you say about the shape of this distribution?

E78. These statistics summarize a set of television ratings from a week without any special programming. Are there any outliers among the 113 ratings?
   Variable N Mean Median TrMean StDev
   Total 113 6.867 6.900 6.596 3.490
   Variable Min Max Q1 Q3
   Total 1.400 20.700 4.550 8.250

E79. The boxplots in Display 2.85 show the life expectancies for the countries of Africa, Europe, and the Middle East. The table shows a few of the summary statistics for each of the three data sets.

   a. From your knowledge of the world, match the boxplots to the correct region.
   b. Match the summary statistics (for Groups A–C) to the correct boxplot (for Regions 1–3).

E80. The National Climatic Data Center records high and low temperatures by state since 1890. Stem-and-leaf plots of the years each state had its lowest temperature and the years each state had its highest temperature are shown in Display 2.86. What do the stems represent? What do the leaves represent? Compare the two distributions with respect to shape, center, spread, and any interesting features.

   E81. A distribution is symmetric with approximately equal mean and median. Is it necessarily the case that about 68% of the values are within one standard deviation of the mean? If yes, explain why. If not, give an example.
E82. Display 2.87 shows two sets of graphs. The first set shows smoothed histograms I–IV for four distributions. The second set shows the corresponding cumulative relative frequency plots, in scrambled order A–D. Match each plot in the first set with its counterpart in the second set.

Distributions
I.  

II.  

III.  

IV.  

Cumulative relative frequency plots
A.  

B.  

C.  

D.  

Display 2.87  Four distributions with different shapes and their cumulative relative frequency plots.

E83. The average number of pedestrian deaths annually for 41 metropolitan areas is given in Display 2.88.

<table>
<thead>
<tr>
<th>Metro Area</th>
<th>Average Annual Deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atlanta</td>
<td>84</td>
</tr>
<tr>
<td>Baltimore</td>
<td>66</td>
</tr>
<tr>
<td>Boston</td>
<td>22</td>
</tr>
<tr>
<td>Charlotte, NC</td>
<td>29</td>
</tr>
<tr>
<td>Chicago</td>
<td>180</td>
</tr>
<tr>
<td>Cincinnati</td>
<td>23</td>
</tr>
<tr>
<td>Cleveland</td>
<td>36</td>
</tr>
<tr>
<td>Columbus, OH</td>
<td>20</td>
</tr>
<tr>
<td>Dallas</td>
<td>76</td>
</tr>
<tr>
<td>Denver</td>
<td>28</td>
</tr>
<tr>
<td>Detroit</td>
<td>107</td>
</tr>
<tr>
<td>Fort Lauderdale</td>
<td>58</td>
</tr>
<tr>
<td>Houston</td>
<td>101</td>
</tr>
<tr>
<td>Indianapolis</td>
<td>24</td>
</tr>
<tr>
<td>Kansas City</td>
<td>27</td>
</tr>
<tr>
<td>Los Angeles</td>
<td>299</td>
</tr>
<tr>
<td>Miami</td>
<td>100</td>
</tr>
<tr>
<td>Milwaukee</td>
<td>19</td>
</tr>
<tr>
<td>Minneapolis</td>
<td>35</td>
</tr>
<tr>
<td>Nassau-Suffolk, NY</td>
<td>80</td>
</tr>
<tr>
<td>Newark, NJ</td>
<td>51</td>
</tr>
<tr>
<td>New Orleans</td>
<td>47</td>
</tr>
<tr>
<td>New York</td>
<td>310</td>
</tr>
<tr>
<td>Norfolk, VA</td>
<td>25</td>
</tr>
<tr>
<td>Orlando, FL</td>
<td>48</td>
</tr>
<tr>
<td>Philadelphia</td>
<td>120</td>
</tr>
<tr>
<td>Phoenix</td>
<td>79</td>
</tr>
<tr>
<td>Pittsburgh</td>
<td>33</td>
</tr>
<tr>
<td>Portland, OR</td>
<td>34</td>
</tr>
<tr>
<td>Riverside, CA</td>
<td>92</td>
</tr>
<tr>
<td>Rochester, NY</td>
<td>17</td>
</tr>
<tr>
<td>Sacramento, CA</td>
<td>37</td>
</tr>
<tr>
<td>Salt Lake City</td>
<td>28</td>
</tr>
<tr>
<td>San Antonio</td>
<td>37</td>
</tr>
<tr>
<td>San Diego</td>
<td>96</td>
</tr>
<tr>
<td>San Francisco</td>
<td>43</td>
</tr>
<tr>
<td>San Jose, CA</td>
<td>33</td>
</tr>
<tr>
<td>Seattle</td>
<td>37</td>
</tr>
<tr>
<td>St. Louis</td>
<td>51</td>
</tr>
<tr>
<td>Tampa</td>
<td>85</td>
</tr>
<tr>
<td>Washington, DC</td>
<td>98</td>
</tr>
</tbody>
</table>

Display 2.88  Average annual pedestrian deaths.  
[Source: Environmental Working Group and the Surface Transportation Policy Project.  
Compiled from National Highway Traffic Safety Administration and U.S. Census data. USA Today, April 9, 1997.]
a. What is the median number of deaths? Write a sentence explaining the meaning of this median.

b. Is any city an outlier in terms of the number of deaths? If so, what is the city, and what are some possible explanations?

c. Make a plot of the data that you think will show the distribution in a useful way. Describe why you chose that plot and what information it gives you about average annual pedestrian deaths.

d. In which situations might giving the death rate be more meaningful than giving the number of deaths?

E84. The side-by-side boxplots in Display 2.89 give the percentage of 4th-grade-age children who are still in school on various continents according to the United Nations. Each case is a country. The four regions marked 1, 2, 3, and 4 are Africa, Asia, Europe, and South/Central America, not necessarily in that order.

a. Which region do you think corresponds to which number?

b. Is the distribution of values for any region skewed left? Skewed right? Symmetrical?

c. Match each dot plot to the corresponding boxplot.

d. In what ways do the boxplots and dot plots give different impressions? Why does this happen? Which type of plot gives a better impression of the distributions?

E85. The first AP Statistics Exam was given in 1997. The distribution of scores received by the 7667 students who took the exam is given in Display 2.91. Compute the mean and standard deviation of the scores.

<table>
<thead>
<tr>
<th>Score</th>
<th>Number of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1205</td>
</tr>
<tr>
<td>4</td>
<td>1696</td>
</tr>
<tr>
<td>3</td>
<td>1873</td>
</tr>
<tr>
<td>2</td>
<td>1513</td>
</tr>
<tr>
<td>1</td>
<td>1380</td>
</tr>
</tbody>
</table>

Display 2.90 shows dot plots of the same data.

Display 2.90 Dot plots of the percentage of 4th-grade-age children still in school in countries of the world, by continent.

c. Match each dot plot to the corresponding boxplot.

d. In what ways do the boxplots and dot plots give different impressions? Why does this happen? Which type of plot gives a better impression of the distributions?
E86. For the countries of Europe, many average life expectancies are approximately the same, as you can see from the stemplot in Display 2.53 on page 69. Use the formulas for the summary statistics of values in a frequency table to compute the mean and standard deviation of the life expectancies for the countries of Europe.

E87. Construct a set of data in which all values are larger than 0, but one standard deviation below the mean is less than 0.

E88. Without computing, what can you say about the standard deviation of this set of values: 4, 4, 4, 4, 4, 4, 4, 4?

E89. In this exercise, you will compare how dividing by \( n \) versus \( n - 1 \) affects the SD for various values of \( n \). So that you don't have to compute the sum of the squared deviations each time, assume that this sum is 400.
   a. Compare the standard deviation that would result from
      i. dividing by 10 versus dividing by 9
      ii. dividing by 100 versus dividing by 99
      iii. dividing by 1000 versus dividing by 999
   b. Does the decision to use \( n \) or \( n - 1 \) in the formula for the standard deviation matter very much if the sample size is large?

E90. If two sets of test scores aren't normally distributed, it's possible to have a larger \( z \)-score on Test II than on Test I yet be in a lower percentile on Test II than on Test I. The computations in this exercise will illustrate this point.
   a. On Test I, a class got these scores: 11, 12, 13, 14, 15, 16, 17, 18, 19, 20. Compute the \( z \)-score and the percentile for the student who got a score of 19.
   b. On Test II, the class got these scores: 1, 1, 1, 1, 1, 1, 18, 19, 20. Compute the \( z \)-score and the percentile for the student who got a score of 18.
   c. Do you think the student who got a score of 19 on Test I or the student who got a score of 18 on Test II did better relative to the rest of the class?

E91. The average income, in dollars, of people in each of the 50 states was computed for 1980 and for 2000. Summary statistics for these two distributions are given in Display 2.92.

<table>
<thead>
<tr>
<th></th>
<th>1980</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$9,725</td>
<td>$28,336</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>1,503</td>
<td>4,413</td>
</tr>
<tr>
<td>Minimum</td>
<td>7,007</td>
<td>21,007</td>
</tr>
<tr>
<td>Lower Quartile</td>
<td>8,420</td>
<td>25,109</td>
</tr>
<tr>
<td>Median</td>
<td>9,764</td>
<td>28,045</td>
</tr>
<tr>
<td>Upper Quartile</td>
<td>10,746</td>
<td>30,871</td>
</tr>
<tr>
<td>Maximum</td>
<td>14,866</td>
<td>41,495</td>
</tr>
</tbody>
</table>

   a. Explain the meaning of $7,007 for the minimum in 1980.
   b. Are any states outliers for either year?
   c. In 2000 the average personal income in Alabama was $23,768, and in 1980 it was $7,836. Did the income in Alabama change much in relation to the other states? Explain your reasoning.

E92. For these comparisons, you will either use the SAT I critical reading scores in Display 2.69 on page 78 or assume that the scores have a normal distribution with mean 505 and standard deviation 111.
   a. Estimate the percentile for an SAT I critical reading score of 425 using the cumulative relative frequency plot. Then find the percentile for a score of 425 using a \( z \)-score. Are the two values close?
   b. Estimate the SAT I critical reading score that falls at the 40th percentile, using the table in Display 2.69. Then find the 40th percentile using a \( z \)-score. Are the two values close?
   c. Estimate the median from the cumulative relative frequency plot. Is this value close to the median you would get by assuming a normal distribution of scores?
d. Estimate the quartiles and the interquartile range using the plot. Find the quartiles and interquartile range assuming a normal distribution of scores.

E93. For 17-year-olds in the United States, blood cholesterol levels in milligrams per deciliter have an approximately normal distribution with mean 176 mg/dL and standard deviation 30 mg/dL. The middle 90% of the cholesterol levels are between what two values?

E94. Display 2.93 shows the distribution of batting averages for all 187 American League baseball players who batted 100 times or more in a recent season. (A batting “average” is the fraction of times that a player hits safely—that is, the hit results in a player advancing to a base—usually reported to three decimal places.)

![Batting Average Distribution](image1)


  a. Do the batting averages appear to be approximately normally distributed?
  b. Approximate the mean and standard deviation of the batting averages from the histogram.
  c. Use your mean and SD from part b to compute an estimate of the percentage of players who batted over .300 (or 300).
  d. Now use the histogram to estimate the percentage of players who batted over .300. Compare to your estimate from part c.

E95. How good are batters in the National League? Display 2.94 shows the distribution of batting averages for all 223 National Leaguers who batted 100 times or more in a recent season.

![Batting Average Distribution](image2)


  a. Approximate the mean and standard deviation of the batting averages from the histogram.
  b. Compare the distributions of batting averages for the two leagues. (See E94 for the American League.) What are the main differences between the two distributions?
  c. A batter hitting .300 in the National League is traded to a team in the American League. What batting average could be expected of him in his new league if he maintains about the same position in the distribution relative to his peers?
AP1. These summary statistics are for the distribution of the populations of the major cities in Brazil.

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>TrMean</th>
<th>StDev</th>
<th>SEMean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>222</td>
<td>381056</td>
<td>191348</td>
<td>261985</td>
<td>820246</td>
<td>55051</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Min</th>
<th>Max</th>
<th>Q1</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>100049</td>
<td>10009231</td>
<td>129542</td>
<td>324323</td>
</tr>
</tbody>
</table>

Which of the following best describes the shape of this distribution?

- A. skewed right without outliers
- B. skewed right with at least one outlier
- C. roughly normal, without outliers
- D. skewed left without outliers
- E. skewed left with at least one outlier

AP2. Which of these lists contains only summary statistics that are sensitive to outliers?

- A. mean, median, and mode
- B. standard deviation, IQR, and range
- C. mean and standard deviation
- D. median and IQR
- E. five-number summary

AP3. This stem-and-leaf plot shows the ages of CEOs of 60 corporations whose annual sales were between $5 million and $350 million. Which of the following is not a correct statement about this distribution?

- A. The distribution is skewed left (towards smaller numbers).
- B. The oldest of the 60 CEOs is 74 years old.
- C. The distribution has no outliers.
- D. The range of the distribution is 42.
- E. The median of the distribution is 50.

AP4. A traveler visits Europe and stays thirty days in thirty different hotels, paying each day with her credit card. The hotels charged a mean price of 50 euros, with a standard deviation of 10 euros. When the charges appear on her credit card statement in the United States, she finds that her bank charged her $1.20 per euro, plus a $5 fee for each transaction. What is the mean and standard deviation of the thirty daily hotel charges in dollars, including the fee?

- A. mean $50, standard deviation $17
- B. mean $60, standard deviation $12
- C. mean $60, standard deviation $17
- D. mean $65, standard deviation $12
- E. mean $65, standard deviation $17

AP5. The scores on a nationally administered test are approximately normally distributed with mean 47.3 and standard deviation 17.3. Approximately what must a student have scored to be in the 95th percentile nationally?

- A. 55
- B. 61
- C. 73
- D. 76
- E. 81

AP6. A particular brand of cereal boxes is labeled “16 oz.” This dot plot shows the actual weights of 100 randomly selected boxes. Which of the following is the best estimate of the standard deviation of these weights?

- A. 0.04 oz.
- B. 0.1 oz.
- C. 0.2 oz.
- D. 0.4 oz.
- E. between 16.0 and 16.2 oz.

AP7. The distribution of the number of points earned by the thousands of contestants in the Game of Pig World Championship has mean 20 and standard deviation 6. What proportion of the contestants earned more than 26 points?
AP8. Anya scored 70 on a statistics test for which the mean was 60 and the standard deviation was 10. She also scored 60 on a chemistry test for which the mean was 50 and the standard deviation was 5. If the scores for both tests were approximately normally distributed, which best describes how Anya did relative to her classmates?

- A. Anya did better on the statistics test than she did on the chemistry test because she scored 10 points higher on the statistics test than on the chemistry test.
- B. Anya did equally well relative to her classmates on each test, because she scored 10 points above the mean on each.
- C. Anya did better on the chemistry test than she did on the statistics test because she scored 2 standard deviations above the mean on the chemistry test and only 1 standard deviation above the mean on the statistics test.
- D. It’s impossible to tell without knowing the number of students in each class.
- E. It’s impossible to tell without knowing the number of points possible on each test.

Investigative Task

AP9. A game invented by three college students involves giving the name of an actor and then trying to connect that actor with actor Kevin Bacon, counting the number of steps needed. For example, Sarah Jessica Parker has a “Bacon number” of 1 because she appeared in the same movie as Kevin Bacon, Footloose (1984). Will Smith has a Bacon number of 2. He has never appeared in a movie with Kevin Bacon; however, he was in Bad Boys II (2003) with Michael Shannon, who was in The Woodsman (2004) with Kevin Bacon. Display 2.95 gives the number of links required to connect each of the 645,957 actors in the Internet Movie Database to Kevin Bacon.

<table>
<thead>
<tr>
<th>Bacon Number</th>
<th>Number of Actors</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1,806</td>
</tr>
<tr>
<td>2</td>
<td>145,024</td>
</tr>
<tr>
<td>3</td>
<td>395,126</td>
</tr>
<tr>
<td>4</td>
<td>95,497</td>
</tr>
<tr>
<td>5</td>
<td>7,451</td>
</tr>
<tr>
<td>6</td>
<td>933</td>
</tr>
<tr>
<td>7</td>
<td>106</td>
</tr>
<tr>
<td>8</td>
<td>13</td>
</tr>
</tbody>
</table>

Display 2.95 Bacon numbers. [Source: www.cs.virginia.edu]

a. How many people have appeared in a movie with Kevin Bacon?

b. Who is the person with Bacon number 0?

It has been questioned whether Kevin Bacon was the best choice for the “center of the Hollywood universe.” A possible challenger is Sean Connery. See Display 2.96.

<table>
<thead>
<tr>
<th>Connery Number</th>
<th>Number of Actors</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2,272</td>
</tr>
<tr>
<td>2</td>
<td>218,560</td>
</tr>
<tr>
<td>3</td>
<td>380,721</td>
</tr>
<tr>
<td>4</td>
<td>40,263</td>
</tr>
<tr>
<td>5</td>
<td>3,537</td>
</tr>
<tr>
<td>6</td>
<td>535</td>
</tr>
<tr>
<td>7</td>
<td>66</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
</tr>
</tbody>
</table>

Display 2.96 Connery numbers.

c. Do you think Kevin Bacon or Sean Connery better deserves the title “Hollywood center”? Make your case using statistical evidence (as always).

d. (For movie fans.) What is Bacon’s Connery number? What is Connery’s Bacon number?
What variables contribute to a college having a high graduation rate? Scatterplots, correlation, and regression are the basic tools used to describe relationships between two quantitative variables.
In Chapter 2, you compared the speeds of predators and nonpredators. Not surprisingly, among mammals meat eaters were usually faster than vegetarians. Some nonpredators, however, such as the horse (48 mi/h) and the elk (45 mi/h), were faster than some predators, such as the dog (39) and the grizzly (30). Because of this variability, comparing the two groups was a matter for statistics; that is, you needed suitable plots and summaries.

The comparison involved a relationship between two variables, one quantitative (speed) and one categorical (predator or not). In this chapter, you’ll learn how to explore and summarize relationships in which both variables are quantitative. The data set on mammals in Display 2.24 on page 43 raises many questions of this sort: Do mammals with longer average longevity also have longer maximum longevity? Is there a relationship between speed and longevity?

The approach to describing distributions in Chapter 2 boiled down to finding shape, center, and spread. For distributions that are approximately normal, two numerical summaries—the mean for center, the standard deviation for spread—tell you basically all you need to know. When comparing two quantitative variables, you can see the shape of the distribution by making a scatterplot. For scatterplots with points that lie in an oval cloud, it turns out once again that two summaries tell you pretty much all you need to know: the regression line and the correlation. The regression line tells about center: What is the equation of the line that best fits the cloud of points? The correlation tells about spread: How spread out are the points around the line?

**In this chapter, you will learn to**

- describe the pattern in a scatterplot, and decide what its *shape* tells you about the relationship between the two variables
- find a regression line through the *center* of a cloud of points to summarize the relationship
- use the correlation as a measure of how *spread* out the points are from this line
- use diagnostic tools to check for information the summaries don’t tell you, and decide what to do with that information
- make shape-changing transformations to re-express a curved relationship so that you can use a line as a summary
In Chapter 1, you explored the relationship between the ages of employees at Westvaco and whether the employees were laid off when the company downsized. There is more to see. In the scatterplot in Display 3.1, for example, each employee is represented by a dot that shows the year the employee was born plotted against the year the employee was hired.

![Scatterplot showing birth year versus hire year for 50 employees in Westvaco Corporation's engineering department.](image)

Display 3.1  *Year of birth versus year of hire* for the 50 employees in Westvaco Corporation's engineering department.

In this scatterplot, you can see a moderate positive association: Employees hired in an earlier year generally were born in an earlier year, and employees hired in a later year generally were born in a later year. This trend is fairly linear. You can visualize a summary line going through the center of the data from lower left to upper right. As you move to the right along this line, the points fan out and cluster less closely around the line.

Sometimes it’s easier to think about people’s ages than about the years they were born. The scatterplot in Display 3.2 shows the ages of the Westvaco employees at the time layoffs began plotted against the year they were hired. This scatterplot shows a moderate negative association: Those people hired in later years generally were younger at the time of the layoffs than people hired in earlier years.

![Scatterplot showing age at layoff versus hire year for 50 employees in Westvaco Corporation's engineering department.](image)

Display 3.2  *Age at layoffs versus year of hire* for the 50 employees in Westvaco Corporation's engineering department.
Interpreting Scatterplots

D1. You will examine Displays 3.1 and 3.2 more closely in these questions and in E7 on page 114.
   a. Why should the two variables plotted in Display 3.1 show a positive association and the two variables plotted in Display 3.2 show a negative association?
   b. Why do all but one of the points in Display 3.1 lie on or below a diagonal line running from the lower left to the upper right?
   c. Is this sentence a reasonable interpretation of Display 3.2? “As time passed, Westvaco tended to hire younger and younger people.”

Describing the Pattern in a Scatterplot

For the distribution of a single quantitative variable, “shape, center, and spread” is a useful summary. For bivariate (two-variable) quantitative data, the summary becomes “shape, trend, and strength.”

You might find it helpful to follow this set of steps as you practice describing scatterplots.

1. Identify the variables and cases. On a scatterplot, each point represents a case, with the x-coordinate equal to the value of one variable and the y-coordinate equal to the value of the other variable. You should describe the scale (units of measurement) and range of each variable.

2. Describe the overall shape of the relationship, paying attention to
   - linearity: Is the pattern linear (scattered about a line) or curved?
   - clusters: Is there just one cluster, or is there more than one?
   - outliers: Are there any striking exceptions to the overall pattern?

3. Describe the trend. If as x gets larger y tends to get larger, there is a positive trend. (The cloud of points tends to slope upward as you go from left to right.) If as x gets larger y tends to get smaller, there is a negative trend. (The cloud of points tends to slope downward as you go from left to right.)

4. Describe the strength of the relationship. If the points cluster closely around an imaginary line or curve, the association is strong. If the points are scattered farther from the line, the association is weak.

   If, as in Display 3.1, the points tend to fan out at one end (a tendency called heteroscedasticity), the relationship varies in strength. If not, it has constant strength.

5. Does the pattern generalize to other cases, or is the relationship an instance of “what you see is all there is”?

6. Are there plausible explanations for the pattern? Is it reasonable to conclude that one variable causes the other? Is there a third or lurking variable that might be causing both?
Example: Dormitory Populations

The plot in Display 3.3 shows, for the 50 states in the United States, the number of people living in college dormitories versus the number of people living in cities, in thousands. Describe the pattern in the plot.

![Plot of Dormitory and Urban Populations](image)

Display 3.3 Number of people living in college dormitories versus number of people living in cities for the 50 states in the United States. [Source: U.S. Census Bureau, 2000 Census of Population and Housing.]

Solution

1. **Variables and cases.** The scatterplot plots dormitory population against urban population, in thousands, for the 50 U.S. states. Dormitory population ranges from near 0 to a high of more than 174,000 in New York. The urban population ranges from near 0 to about 17 million in Texas and New York and 32 million in California.

2. **Shape.** While most states follow a linear trend, the three states with the largest urban population suggest curvature in the plot because, for those states, the number of people living in dormitories is proportionately lower than in the smaller states. California can be considered an outlier with respect to its urban population, which is much larger than that of other states. It is also an outlier with respect to the overall pattern, because it lies far below the generally linear trend.

3. **Trend.** The trend is positive—states with larger urban populations tend to have larger dormitory populations, and states with smaller urban populations tend to have smaller dormitory populations.

4. **Strength.** The relationship varies in strength. For the states with the smallest urban populations, the points cluster rather closely around a line. For the states with the largest urban populations, the points are scattered farther from the line. Overall, the strength of the relationship is moderate.

5. **Generalization.** The 50 states aren’t a sample from a larger population of cases, so the relationship here does not generalize to other cases. Because both variables tend to change rather slowly, however, we can expect the relationship in Display 3.3 to be similar to that of other years.
6. **Explanation.** It is tempting to attribute the positive relationship to the idea that cities attract colleges. (Just pick a large city nearby and see how many colleges you can name that are located there.) The main reason for the positive relationship, however, is not nearly so interesting: Both variables are related to a state’s population. The more people in a state, the more people live in dormitories and the more people live in cities. (There’s a moral here: Interpreting association can be tricky, in part because the two variables you see in a plot often will be related to some lurking variable that you don’t see.)

---

**DISCUSSION**

**Describing the Pattern in a Scatterplot**

D2. Display 3.4 is derived from the data for Display 3.3 by converting the variables to the proportion of a state’s population living in college dormitories (given as the number living in dorms per 1000 state residents) and the proportion of the state’s population living in cities.

a. Follow steps 1–6 in the previous example to describe what you see in this new plot.

b. When you go from totals (Display 3.3) to proportion of total population (Display 3.4), the relationship changes from positive to negative and becomes weaker. Give an explanation for the differences in these two plots.

![Display 3.4](image-url)

**Display 3.4** The proportion of people living in college dormitories versus the proportion of people living in cities for the 50 U.S. states.

---

**Summary 3.1: Scatterplots**

A scatterplot shows the relationship between two quantitative variables. Each case is a point, with the x-coordinate equal to the value of one variable and the y-coordinate equal to the value of the other variable.

In describing a scatterplot, be sure to cover all of the following:

- cases and variables (What exactly does each point represent?)
- shape (linear or curved, clusters, outliers)
• trend (positive, negative, or none)
• strength (strong, moderate, or weak; constant or varying)
• generalization (Does the pattern generalize?)
• explanation (Is there an explanation for the pattern?)

When asked to **compare** two scatterplots, don't just describe each separately. Be sure to describe how their shapes, trends, and strengths are similar and how they differ.

**Practice**

**Describing the Pattern in a Scatterplot**

**P1. Growing kids.** This table gives median heights of boys at ages 2, 3, 4, 5, 6, and 7 yr.

<table>
<thead>
<tr>
<th>Age (yr)</th>
<th>Height (in.)</th>
<th>Age (yr)</th>
<th>Height (in.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>35.8</td>
<td>5</td>
<td>44.2</td>
</tr>
<tr>
<td>3</td>
<td>39.1</td>
<td>6</td>
<td>46.8</td>
</tr>
<tr>
<td>4</td>
<td>41.4</td>
<td>7</td>
<td>49.6</td>
</tr>
</tbody>
</table>

a. **Scatterplot.** Plot height versus age; that is, put height on the y-axis and age on the x-axis.

b. **Shape, trend, and strength.** Describe the shape, trend, and strength of the relationship.

c. **Generalization.** Would you expect these data to allow you to make good predictions of the median height of 8-year-olds? Of 50-year-olds?

d. **Explanation.** It doesn't quite fit to say that age "causes" height, but there is still an underlying cause-and-effect relationship. How would you describe it?

**P2. Late planes and lost bags.** A great way to cap off a long day of travel is to have your plane arrive late and then find that the airline has lost your luggage. As Display 3.5 shows, some airlines handle baggage better than others.

a. Which airline has the worst record for mishandled baggage? For being on time?

b. Where on the plot would you find the airline with the best on-time record and the best mishandled-baggage rate? Which airlines are best in both categories?

c. Determine whether this statement is true or false, and explain your answer: American had a mishandled-baggage rate that was more than twice the rate of Southwest.

d. Is there a positive or a negative relationship between the on-time percentage and the rate of mishandled baggage? Is it strong or weak?

e. Would you expect the relationship in this plot to generalize to some larger population of commercial airlines? Why or why not? Would you expect the relationship in this plot to be roughly the same for data from 10 years ago? For next year?
Exercises

E1. For each of the lettered scatterplots in Display 3.6, give the trend (positive or negative), strength (strong, moderate, or weak), and shape (linear or curved). Which plots show varying strength?

Display 3.6 Eight scatterplots with various distributions.

E2. For each set of cases and variables, tell whether you expect the relationship to be (i) positive or negative and (ii) strong, moderate, or weak.

<table>
<thead>
<tr>
<th>Cases</th>
<th>Variable 1</th>
<th>Variable 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Hens' eggs</td>
<td>Length</td>
<td>Width (diameter of cross-section)</td>
</tr>
<tr>
<td>b. High school seniors</td>
<td>SAT I math score</td>
<td>SAT I critical reading score</td>
</tr>
<tr>
<td>c. Trees</td>
<td>Age</td>
<td>Number of rings</td>
</tr>
<tr>
<td>d. People</td>
<td>Age</td>
<td>Body flexibility</td>
</tr>
<tr>
<td>e. U.S. states</td>
<td>Population</td>
<td>Number of representatives in Congress</td>
</tr>
<tr>
<td>f. Countries of the United Nations</td>
<td>Land area</td>
<td>Population</td>
</tr>
<tr>
<td>g. Olympic games</td>
<td>Year</td>
<td>Winning time in the women's 100-meter race</td>
</tr>
</tbody>
</table>

LaTasha Colander crosses the finish line of the women's 100-meter dash final at the 2004 U.S. Olympic Team Track and Field Trials.
E3. Match each set of cases and variables (A–D) with the short summary (I–IV) of its scatterplot.

<table>
<thead>
<tr>
<th>Cases</th>
<th>Variable 1</th>
<th>Variable 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Earth</td>
<td>Year</td>
<td>Human population</td>
</tr>
<tr>
<td>B. U.S. states</td>
<td>Area in square miles</td>
<td>People per square mile</td>
</tr>
<tr>
<td>C. U.S. states</td>
<td>Population</td>
<td>Number of doctors</td>
</tr>
<tr>
<td>D. Cars in the United States</td>
<td>Weight</td>
<td>Gas mileage</td>
</tr>
</tbody>
</table>

I. strong negative relationship, somewhat curved
II. strong, curved positive relationship
III. moderate, roughly linear, positive relationship
IV. moderate negative relationship

E4. SAT I math scores. In 2005, the average SAT I math score across the United States was 520. North Dakota students averaged 605, Illinois students averaged 606, and students from the nearby state of Iowa did even better, averaging 608. Why do states from the Midwest do so well? It is easy to jump to a false conclusion, but the scatterplot in Display 3.7 can help you find a reasonable explanation.

a. Estimate the percentage of students in Iowa and in Illinois who took the SAT I. New York had the highest percentage of students who took the SAT I. Estimate that percentage and the average SAT I math score for students in that state.

b. Describe the shape of the plot. Do you see any clusters? Are there any outliers? Is the relationship linear or curved? Is the overall trend positive or negative? What is the strength of the relationship?

c. Is the distribution of the percentage of students taking the SAT I bimodal? Explain how the scatterplot shows this. Is the distribution of SAT I math scores bimodal?

d. The cases used in this plot are the 50 U.S. states in 2005. Would you expect the pattern to generalize to some other set of cases? Why or why not?

e. Suggest an explanation for the trend. (Hint: The SAT is administered from Princeton, New Jersey. An alternative exam, the ACT, is administered from Iowa. Many colleges and universities in the Midwest either prefer the ACT or at least accept it in place of the SAT, whereas colleges in the eastern states tend to prefer the SAT.) Is there anything in the data that you can use to help you decide whether your explanation is correct?
E5. Each of the 51 cases plotted on the scatterplots in Display 3.8 is a top-rated university. The $y$-coordinate of a point tells the graduation rate, and the $x$-coordinate tells the value of some other quantitative variable—the percentage of alumni who gave that year, the student/faculty ratio, the 75th percentile of the SAT scores (math plus critical reading) for a recent entering class, and the percentage of incoming students who ranked in the top 10% of their high school graduating class.

Display 3.8 Scatterplots showing the relationship between graduation rate and four other variables for 51 top-rated universities. [Source: U.S. News and World Report, 2000.]

a. Compare the shapes of the four plots.
   i. Which plots show a linear shape? Which show a curved shape?
   ii. Which plots show just one cluster? Which show more than one?
   iii. Which plots have outliers?

b. Compare the trends of the relationships:
   Which plots show a positive trend? A negative trend? No trend?

c. Compare the strengths of the relationships: Which variables give more precise predictions of the graduation rate? Which variable is almost useless for predicting graduation rate?

d. Generalization. The cases in these plots are the 51 universities that happened to come out at the top of one particular rating scheme. Do you think the complete set of all U.S. universities would show pretty much the same relationships? Why or why not?

e. Explanation. Consider the two variables with the strongest relationship to graduation rates. Offer an explanation for the strength of these particular relationships. In what ways, if any, can you use the data to help you decide whether your explanation is in fact correct?

E6. Hat size. What does hat size really measure? A group of students investigated this question by collecting a sample of hats. They recorded the size of the hat and then measured the circumference, the major axis (the length across the opening in the long direction), and the minor axis. (See Display 3.9 on the next page. Hat sizes have been changed to decimals; all other measurements are in inches.) Is hat size most closely related to circumference, major axis, or minor axis? Is hat size most closely related to circumference, major axis, or minor axis? Answer this question by making appropriate plots and describing the patterns in those plots.
Display 3.9  Hat sizes, with circumference and axes in inches. [Source: Roger Johnson, Carleton College, data from student project.]

<table>
<thead>
<tr>
<th>Hat Size</th>
<th>Circumference</th>
<th>Major Axis</th>
<th>Minor Axis</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.625</td>
<td>20.00</td>
<td>7.00</td>
<td>5.75</td>
</tr>
<tr>
<td>6.750</td>
<td>20.75</td>
<td>7.25</td>
<td>6.00</td>
</tr>
<tr>
<td>6.875</td>
<td>20.50</td>
<td>7.50</td>
<td>6.00</td>
</tr>
<tr>
<td>6.875</td>
<td>20.75</td>
<td>7.25</td>
<td>6.00</td>
</tr>
<tr>
<td>6.875</td>
<td>20.75</td>
<td>7.50</td>
<td>6.00</td>
</tr>
<tr>
<td>6.875</td>
<td>21.50</td>
<td>7.25</td>
<td>6.25</td>
</tr>
<tr>
<td>7.000</td>
<td>21.25</td>
<td>7.50</td>
<td>6.00</td>
</tr>
<tr>
<td>7.000</td>
<td>21.00</td>
<td>7.50</td>
<td>6.00</td>
</tr>
<tr>
<td>7.000</td>
<td>21.00</td>
<td>7.50</td>
<td>6.25</td>
</tr>
<tr>
<td>7.000</td>
<td>21.75</td>
<td>7.50</td>
<td>6.25</td>
</tr>
<tr>
<td>7.125</td>
<td>21.50</td>
<td>7.75</td>
<td>6.25</td>
</tr>
<tr>
<td>7.125</td>
<td>21.75</td>
<td>7.75</td>
<td>6.50</td>
</tr>
<tr>
<td>7.125</td>
<td>21.50</td>
<td>7.75</td>
<td>6.25</td>
</tr>
<tr>
<td>7.125</td>
<td>22.25</td>
<td>7.75</td>
<td>6.25</td>
</tr>
<tr>
<td>7.250</td>
<td>22.00</td>
<td>7.75</td>
<td>6.25</td>
</tr>
<tr>
<td>7.250</td>
<td>22.50</td>
<td>7.75</td>
<td>6.50</td>
</tr>
<tr>
<td>7.375</td>
<td>22.25</td>
<td>7.75</td>
<td>6.50</td>
</tr>
<tr>
<td>7.375</td>
<td>22.25</td>
<td>8.00</td>
<td>6.50</td>
</tr>
<tr>
<td>7.375</td>
<td>22.50</td>
<td>8.00</td>
<td>6.50</td>
</tr>
<tr>
<td>7.375</td>
<td>22.75</td>
<td>8.00</td>
<td>6.50</td>
</tr>
<tr>
<td>7.375</td>
<td>23.00</td>
<td>8.00</td>
<td>6.50</td>
</tr>
<tr>
<td>7.500</td>
<td>22.75</td>
<td>8.00</td>
<td>6.50</td>
</tr>
<tr>
<td>7.500</td>
<td>22.50</td>
<td>8.00</td>
<td>6.50</td>
</tr>
<tr>
<td>7.625</td>
<td>23.00</td>
<td>8.25</td>
<td>6.50</td>
</tr>
<tr>
<td>7.625</td>
<td>23.00</td>
<td>8.25</td>
<td>6.50</td>
</tr>
<tr>
<td>7.625</td>
<td>23.25</td>
<td>8.25</td>
<td>6.75</td>
</tr>
</tbody>
</table>

E7. Westvaco, revisited. To determine whether Westvaco discriminated by age in laying off employees, you could investigate whether it might have discriminated in hiring. Display 3.10 shows the age at hire plotted against the year the person was hired.

a. Describe the pattern in the plot, following the six-step model.

b. Does this plot provide evidence that Westvaco discriminated by age in hiring?

c. Display 3.11 shows the year of birth of the Westvaco employees plotted against the year they were hired. Open circles represent employees laid off, and solid circles represent employees kept. Does this scatterplot suggest a reason why older employees tended to be laid off more frequently?

E8. Passenger aircraft. Airplanes vary in their size, speed, average flight length, and cost of operation. You can probably guess that larger planes use more fuel per hour and cost more to operate than smaller planes, but the shapes of the relationships are less obvious. Display 3.12 lists data on the 33 most commonly used passenger airplanes in the United States. The variables are the number of seats, average cargo payload in tons, airborne speed in miles per hour, flight length in miles, fuel consumption in gallons per hour, and operating cost per hour in dollars.

#### a. cost per hour

i. Make scatterplots with cost per hour on the y-axis to explore this variable's dependence on the other variables. Report your most interesting findings. Here are examples of some questions you could investigate: For which variable is the relationship to the cost per hour strongest? Is there any one airplane whose cost per hour, in relation to other variables, makes it an outlier?
ii. Do your results mean that larger planes are less efficient? Define your own variable, and plot it against other variables to judge the relative efficiency of the larger planes.

b. flight length
   i. Make scatterplots with length of flight on the x-axis to explore this variable's relationship to the other variables. Report your most interesting findings. Here is an example of a question you could investigate: Which variable, cargo or number of seats, shows a stronger relationship to flight length? Propose a reasonable explanation for why this should be so.
   ii. Do planes with a longer flight length tend to use less fuel per mile than planes with a shorter flight length?

c. speed, seats, and cargo
   i. Make scatterplots to explore the relationships between the variables speed, seats, and cargo. Report your most interesting findings. Here are some examples of questions you could investigate: For which variable, cargo or number of seats, is the relationship to speed more obviously curved? Explain why that should be the case. Which plane is unusually slow for the amount of cargo it carries? Which plane is unusually slow for the number of seats it has?
   ii. The plot of cargo against seats has two parts: a flat stretch on the left and a fan on the right. Explain, in the language of airplanes, seats, and cargo, what each of the two patterns tells you.

---

3.2 Getting a Line on the Pattern

In this section, you will learn how to use a regression line to summarize the relationship between two quantitative variables. This section deals first with the simple situation in which all the data points lie close to a line. In practice, however, data points are often more scattered. The second part of this section shows how to choose a summary line when your data points form an oval cloud. To begin, you will review the properties of a linear equation.

Lines as Summaries

You've seen the equation of a line, \( y = \text{slope} \cdot x + \text{y-intercept} \), so the review here will be brief. Linear relationships have the important property that for any two points \((x_1, y_1)\) and \((x_2, y_2)\) on the line, the ratio

\[
\frac{\text{rise}}{\text{run}} = \frac{\text{change in} \ y}{\text{change in} \ x} = \frac{y_2 - y_1}{x_2 - x_1}
\]

is a constant. This ratio is the slope of the line. The rise and run are illustrated in Display 3.13, where the slope is the ratio of the two sides of the right triangle. This ratio is the same for any two points on the line because all the triangles formed are similar.
How thick is a single sheet of your book? One sheet alone is too thin to measure directly with a ruler, but you could measure the thickness of 100 sheets together, then divide by 100. This method would give you an estimate of the thickness but no information about how much your estimate is likely to vary from the true thickness. The approach in the next activity lets you judge precision as well as thickness.

**ACTIVITY 3.2a**

**Pinching Pages**

**What you’ll need:** a ruler with a millimeter scale, a copy of your textbook

1. Pinch together the front cover and first 50 sheets of your textbook. Then measure and record the thickness to the nearest millimeter.
2. Repeat for the front cover plus 100, 150, 200, and 250 sheets.
3. Plot your data on a scatterplot, with number of sheets on the horizontal scale and total thickness on the vertical scale.
4. Does the plot look linear? Should it? Discuss why or why not, and make your measurements again if necessary. On the plot, place a straight line that best fits the cloud of points.
5. Find the slope and y-intercept of your line. What does the y-intercept tell you? What does the slope tell you? What is your estimate of the thickness of a sheet?
6. Use the information in your graph to discuss how much your estimate in step 5 is likely to vary from the true thickness.
7. How would your line have changed if you hadn’t included the front cover?
INTERPRETING SLOPE

D3. Decide what the variables $y$ and $x$ represent in these situations.

a. Suppose regular unleaded gasoline costs $2.60 per gallon. The number 2.60 is the slope of the line you get if you plot $y$ versus $x$.

b. Suppose your car averages 30 mi/gal. The number 30 is the slope of a line fitted to a scatterplot of $y$ versus $x$.

c. “A pint’s a pound the world around.” This slogan summarizes the slope of a line fitted to a scatterplot of $y$ versus $x$ for various quantities of various kinds of liquids.

The next example illustrates how to find the equation of a line when you know two points that fall on the line.

Example: Minimum Wage

In an effort to keep wages of hourly workers at a level that allows some possibility of making a decent living, the United States government establishes a minimum hourly wage rate. The scatterplot in Display 3.14 shows the minimum wage (in dollars) for every five years from 1960 through 2005. The line on the plot is the least squares regression line, which you will learn about later in this section.

Estimate the slope of the line. What does the slope tell you? Estimate the equation of the line.

![Display 3.14 Minimum wages at five-year intervals, 1960 through 2005.](image)

Solution

In theory, you can find the slope from any two points on the line. Here, however, you have to estimate the coordinates from the graph. In such cases, you usually can produce a better estimate of the slope by choosing two points that are far apart. For this plot, choosing the points on the line for the years 1960 and 2000 works well. Approximate points are $(1960, 0.80)$ for $(x_1, y_1)$ and $(2000, 4.80)$ for $(x_2, y_2)$. The estimated slope is

$$
slope = \frac{y_2 - y_1}{x_2 - x_1} = \frac{4.80 - 0.80}{2000 - 1960} = \frac{4.0}{40} = 0.10
$$
The slope tells you that the minimum wage increased by about $0.10 per year over the 45-yr period 1960 through 2005.

You can write the equation of a line in terms of its slope and $y$-intercept.

$$y = \text{slope} \cdot x + \text{$y$-intercept}$$

You have the slope, but you cannot read the $y$-intercept from the plot. To find it, use the slope and a point such as (1960, 0.80) to solve for the $y$-intercept.

$$y = \text{slope} \cdot x + \text{$y$-intercept}$$

$$0.80 = 0.10(1960) + \text{$y$-intercept}$$

$$\text{$y$-intercept} = 0.80 - 196 = -195.20$$

The equation of the line using these approximate values is

$$y = 0.10x - 195.20$$

In statistics this equation usually is written with the intercept first, becoming

$$y = -195.20 + 0.10x$$

---

**Lines as Summaries**

D4. The Consumer Price Index (CPI) is a measure of the prices paid by urban consumers for a selected group of goods and services (called a “market basket”) thought to be typical of urban households. The CPI often is used to adjust salaries, rents, and other segments of the economy. The CPI is, itself, a statistical estimate based on a number of large-scale surveys conducted by agencies of the federal government. Display 3.15 shows the CPI for every five years from 1970 to 2005. Fit a line to these data by eye, and use two points to estimate its slope. What is the annual rate of increase for the CPI over this 35-yr period? What is the equation of your line?
Using Lines for Prediction

There are two main reasons why you would want to fit a line to a set of data:

• to find a summary, or model, that describes the relationship between the two variables
• to use the line to predict the value of \( y \) when you know the value of \( x \). In cases where it makes sense to do this, the variable on the \( x \)-axis is called the predictor or explanatory variable, and the variable on the \( y \)-axis is called the predicted or response variable.

In the previous example, the equation

\[
y = -195.20 + 0.10x
\]

models the rise in the minimum wage for the years 1960 through 2005. Knowing this equation enables you to make a general statement about the minimum wage throughout these years: “The minimum wage went up roughly $0.10 per year.”

You might instead want to use the line to predict the minimum wage in one of the years for which no amount is given or for years before 1960 or after 2005.

**Example: Predicting the Minimum Wage**

Use the equation \( y = -195.20 + 0.10x \) to predict the minimum wage in the years 2003 and 1950.

**Solution**

The predicted minimum wage for 2003 is

\[
y = -195.20 + 0.10x = -195.20 + 0.10(2003) = 5.10
\]

Assuming the linear trend continues back to earlier years, the predicted minimum wage for 1950 is

\[
y = -195.20 + 0.10x = -195.20 + 0.10(1950) = -0.20
\]

The predicted minimum wage for 2003 is very close to the actual minimum wage of $5.15 per hour. But the actual minimum wage in 1950 was $0.75 per hour, not a negative number! As you can see, making the assumption that the linear trend continues can be risky. This type of prediction, making a prediction when the value of \( x \) falls outside the range of the actual data, is called extrapolation.

Interpolation—making a prediction when the value of \( x \) falls inside the range of the data, as does 2003—is safer.

Suppose you know the value of \( x \) and use a line to predict the corresponding value of \( y \). You know that your prediction for \( y \) won’t be exact, but you hope that the error will be small. The prediction error is the difference between the observed value of \( y \) and the predicted value of \( y \), or \( \hat{y} \). You usually don’t know what that error is. If you did, you wouldn’t need to use the line to predict the value of \( y \). You do, however, know the errors for the points used to construct the line. These differences are called residuals:

\[
\text{residual} = \text{observed value of } y - \text{predicted value of } y = y - \hat{y}
\]
The geometric interpretation of the residual is shown in Display 3.16. A residual is the signed vertical distance from an observed data point to the regression line. The residual is positive if the point is above the line and negative if the point is below the line.

Display 3.16  Residual = $y - \hat{y}$

**Example: Finding Residuals**

Display 3.17 shows the mean net income (after expenses and before taxes, in thousands of dollars) for doctors who were board-certified in family practice and working during the years 1990–1998 and 2001.

<table>
<thead>
<tr>
<th>Year</th>
<th>Mean Net Income (thousand $)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>103</td>
</tr>
<tr>
<td>1991</td>
<td>112</td>
</tr>
<tr>
<td>1992</td>
<td>114</td>
</tr>
<tr>
<td>1993</td>
<td>117</td>
</tr>
<tr>
<td>1994</td>
<td>121</td>
</tr>
<tr>
<td>1995</td>
<td>131</td>
</tr>
<tr>
<td>1996</td>
<td>139</td>
</tr>
<tr>
<td>1997</td>
<td>141</td>
</tr>
<tr>
<td>1998</td>
<td>143</td>
</tr>
<tr>
<td>2001</td>
<td>145</td>
</tr>
</tbody>
</table>


The equation of the fitted line is $\hat{y} = -8300.6 + 4.2248x$, where $x$ is the year and $\hat{y}$ is the income in thousands of dollars.

Graph the fitted line with the data points. What is the residual for the year 1996?

**Solution**

You can use a graphing calculator to graph a scatterplot with a summary line. [See Calculator Note 3B.]
The actual net income value for 1996 was $139,000. Using the equation of the fitted line, the prediction for 1996 is

$$\hat{y} = -8300.6 + 4.2248x = -8300.6 + 4.2248(1996) = 132.1008,$$

or $132,110$. You also can use your calculator to calculate a predicted value quickly. [See Calculator Note 3C.]

To find the residual, subtract the predicted value from the observed value:

$$y - \hat{y} = 139 - 132.1008 = 6.5992$$

or about $6899$. The residual is positive because the observed value is higher than the predicted value. That is, the point lies above the line.

You can use a calculator to calculate residuals for all points in a data set simultaneously. [See Calculator Note 3D.]

**Discussion**

**Using Lines for Prediction**

D5. Test how well you understand residuals.

a. If a residual is large and negative, where is the point located with respect to the line? Draw a diagram to illustrate. What does it mean if the residual is 0?

b. If someone said that they had fit a line to a set of data points and all their residuals were positive, what would you say to them?

c. Interpret the $y$-intercept of the regression line in the previous example. Does this make sense?
D6. What do you think of the arithmetic and the reasoning in this passage from Mark Twain’s Life on the Mississippi?

In the space of one hundred and seventy-six years the Lower Mississippi has shortened itself two hundred and forty-two miles. That is an average of a trifle over one mile and a third per year. Therefore, any calm person, who is not blind or idiotic, can see that in the Old Oölitic Silurian Period, just a million years ago next November, the Lower Mississippi River was upwards of one million three hundred thousand miles long, and stuck out over the Gulf of Mexico like a fishing rod.

And by the same token any person can see that seven hundred and forty-two years from now the Lower Mississippi will be only a mile and three quarters long, and Cairo and New Orleans will have joined their streets together, and be plodding comfortably along under a single mayor and a mutual board of aldermen. There is something fascinating about science. One gets such wholesale returns of conjecture out of such a trifling investment of fact. [Source: James R. Osgood and Company, 1883, p. 208.]

Given that the Mississippi/Missouri river system was about 3710 mi long in the year 2000, write an equation that Twain would say gives the length of the river in terms of the year.

**Least Squares Regression Lines**

The general approach to fitting lines to data is called the method of least squares. The method was invented about 200 years ago by Carl Friedrich Gauss (1777–1855), Adrien-Marie Legendre (1752–1833), and Robert Adrain (1775–1843), who were working independently of one another in Germany, France, and Ireland, respectively.

The least squares regression line, also called least squares line or regression line, for a set of data points \((x, y)\) is the line for which the sum of squared errors (residuals), or SSE, is as small as possible.

\[
\text{SSE} = \sum (\text{residual})^2 = \sum (y - \hat{y})^2
\]

**Example: Regression Line for the Passenger Jets**

This table shows cost per hour versus number of seats for three models of passenger jets from the data in Display 3.12 on page 115. (Some of the values have been rounded.)

<table>
<thead>
<tr>
<th>Aircraft</th>
<th>Number of Seats</th>
<th>Cost ($/h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERJ-145</td>
<td>50</td>
<td>1100</td>
</tr>
<tr>
<td>DC-9</td>
<td>100</td>
<td>2100</td>
</tr>
<tr>
<td>MD-90</td>
<td>150</td>
<td>2700</td>
</tr>
<tr>
<td>Mean</td>
<td>100</td>
<td>1967</td>
</tr>
</tbody>
</table>
Which of these two equations gives the least squares regression line for predicting cost from number of seats?

\[ \hat{y} = 367 + 16x \]
\[ \hat{y} = 300 + 16x \]

**Solution**

The least squares regression line minimizes the sum of the squared errors, SSE, so the equation with the smaller SSE must be the equation of the regression line.

For the equation \( \hat{y} = 367 + 16x \):

\[
\begin{array}{cccccc}
 x & y & \hat{y} & y - \hat{y} & (y - \hat{y})^2 \\
 50 & 1100 & 1167 & -67 & 4,489 \\
 100 & 2100 & 1967 & 133 & 17,689 \\
 150 & 2700 & 2767 & -67 & 4,489 \\
 \hline 
 & & & & \text{SSE} = 26,667 \\
\end{array}
\]

For the equation \( \hat{y} = 300 + 16x \):

\[
\begin{array}{cccccc}
 x & y & \hat{y} & y - \hat{y} & (y - \hat{y})^2 \\
 50 & 1100 & 1100 & 0 & 0 \\
 100 & 2100 & 1900 & 200 & 40,000 \\
 150 & 2700 & 2700 & 0 & 0 \\
 \hline 
 & & & & \text{SSE} = 40,000 \\
\end{array}
\]

[You can use your calculator to calculate the SSE quickly. See **Calculator Note 3E**.]

The first equation has the smaller SSE, so it must be the equation of the least squares regression line. Note that for this line, except for rounding error, the sum of the residuals, \( \sum (y - \hat{y}) \), is equal to 0. This is always the case for the least squares regression line, but it can be true for other lines, too.

[You can use a calculator program to visually explore the least squares regression line and SSE. See **Calculator Note 3F**.]

In addition to making the sum of the squared errors as small as possible, the least squares regression line has some other properties, given in the box on the next page.
3.2 Getting a Line on the Pattern

Properties of the Least Squares Regression Line

The fact that the sum of squared errors, or SSE, is as small as possible means that, for the least squares regression line, these properties also hold:

- The sum (and mean) of the residuals is 0.
- The line contains the point of averages, \((\bar{x}, \bar{y})\).
- The standard deviation of the residuals is smaller than for any other line that goes through the point \((\bar{x}, \bar{y})\).
- The line has slope \(b_1\), where

\[
b_1 = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}
\]

There are some appealing mathematical relationships among these properties, which, taken together, show that the line through the point of averages \((\bar{x}, \bar{y})\) having slope \(b_1\) does, in fact, minimize the sum of the squared errors. This gives you a way to find the equation of the least squares line: \(\hat{y} = b_0 + b_1x\). First, compute the slope \(b_1\) using the formula in the box. Then find the \(y\)-intercept, \(b_0\), using the point \((\bar{x}, \bar{y})\) and solving the equation \(\bar{y} = b_0 + b_1\bar{x}\) for \(b_0\), \(b_0 = \bar{y} - b_1\bar{x}\).

Example: Least Squares Line for the Passenger Jets

Find the least squares line for the passenger jets data given in the previous example.

Solution

Finding the line requires three main steps: Find the point of averages \((\bar{x}, \bar{y})\), find the slope \(b_1\), and use the point and slope to find the \(y\)-intercept \(b_0\).

A convenient way to organize the computations is to work from a table.

<table>
<thead>
<tr>
<th>Aircraft</th>
<th>Seats, (x)</th>
<th>Cost ($/h), (y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERJ-145</td>
<td>50</td>
<td>1100</td>
</tr>
<tr>
<td>DC-9</td>
<td>100</td>
<td>2100</td>
</tr>
<tr>
<td>MD-90</td>
<td>150</td>
<td>2700</td>
</tr>
<tr>
<td>Sum</td>
<td>300</td>
<td>5900</td>
</tr>
<tr>
<td>Mean</td>
<td>100</td>
<td>1966.6</td>
</tr>
</tbody>
</table>

Point of averages: The point of averages \((\bar{x}, \bar{y})\) is \((100, 1966.6)\), and the least squares regression line passes through this point.
Slope: To compute the slope, first create two new columns for deviations from the mean, one for $x - \bar{x}$ and the other for $y - \bar{y}$:

<table>
<thead>
<tr>
<th>Aircraft</th>
<th>Seats, $x$</th>
<th>Cost ($/h), y$</th>
<th>$x - \bar{x}$</th>
<th>$y - \bar{y}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERJ-145</td>
<td>50</td>
<td>1100</td>
<td>-50</td>
<td>-866.6</td>
</tr>
<tr>
<td>DC-9</td>
<td>100</td>
<td>2100</td>
<td>0</td>
<td>133.3</td>
</tr>
<tr>
<td>MD-90</td>
<td>150</td>
<td>2700</td>
<td>50</td>
<td>733.3</td>
</tr>
<tr>
<td>Sum</td>
<td>300</td>
<td>5900</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mean</td>
<td>100</td>
<td>1966.6</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Now create two more columns, one for $(x - \bar{x}) \cdot (y - \bar{y})$ and the other for $(x - \bar{x})^2$:

<table>
<thead>
<tr>
<th>Aircraft</th>
<th>Seats, $x$</th>
<th>Cost ($/h), y$</th>
<th>$x - \bar{x}$</th>
<th>$y - \bar{y}$</th>
<th>$(x - \bar{x}) \cdot (y - \bar{y})$</th>
<th>$(x - \bar{x})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERJ-145</td>
<td>50</td>
<td>1100</td>
<td>-50</td>
<td>-866.6</td>
<td>4333.3</td>
<td>2500</td>
</tr>
<tr>
<td>DC-9</td>
<td>100</td>
<td>2100</td>
<td>0</td>
<td>133.3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MD-90</td>
<td>150</td>
<td>2700</td>
<td>50</td>
<td>733.3</td>
<td>3666.6</td>
<td>2500</td>
</tr>
<tr>
<td>Sum</td>
<td>300</td>
<td>5900</td>
<td>0</td>
<td>0</td>
<td>80000</td>
<td>5000</td>
</tr>
<tr>
<td>Mean</td>
<td>100</td>
<td>1966.6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The ratio of the sums of the last two columns gives the slope

$$b_1 = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2} = \frac{80000}{5000} = 16$$

$y$-intercept: Now that you have a point on the line, $(100, 1966.6)$, and the slope, 16, you can find the $y$-intercept from the equation

$$b_0 = \bar{y} - b_1\bar{x}$$

$$= 1966.6 - 16(100)$$

$$= 366.6$$

This agrees with what you found in the previous example. That is, the equation of the least squares regression line (with rounded $y$-intercept) is

$$\hat{y} = 367 + 16x$$

[You also can use your calculator to find the equation of the least squares line. See Calculator Note 3G.]

**DISCUSSION**

**Least Squares Regression Lines**

D7. You might have wondered why statisticians don’t fit a regression line by minimizing the sum of the absolute values of the residuals, $\sum |y - \hat{y}|$, rather than the sum of the squares of the residuals. Here you will learn one reason why.
a. Plot the points in the table and the line $y = 1 + x$. Explain why this is the line that best fits these points. Compute the sum of the absolute values of the residuals.

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

b. Draw another line that passes between the two points at $x = 0$ and also passes between the two points at $x = 2$. Compute the sum of the absolute values of the residuals for this line and compare it to your sum from part a.

c. Draw yet another line that passes between the two points at $x = 0$ and also passes between the two points at $x = 2$. Find the sum of the absolute values of the residuals for this line and compare it to your sums from parts a and b.

d. Draw a line that does not pass between the two points at $x = 0$. Find the sum of the absolute values of the residuals for this line and compare it to your sums from parts a and b.

e. Now find the least squares regression line and compute the sum of the squared residuals. Compute the sum of the squared residuals for your lines in parts b, c, and d. What can you conclude?

f. Find the standard deviation of the residuals for the least squares regression line and for the lines in parts b, c, and d. What can you conclude?

**Reading Computer Output**

When you are working with real data, the best way to get the least squares line is by computer or calculator. Display 3.18 shows typical computer output for the minimum wage data in Display 3.14 on page 118.

```
Dependent variable is: MinWage
No Selector
R squared = 98.3%  R squared (adjusted) = 98.1%
s = 0.2105 with 10 - 2 = 8 degrees of freedom

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>21.0017</td>
<td>1</td>
<td>21.0017</td>
<td>474</td>
</tr>
<tr>
<td>Residual</td>
<td>0.354545</td>
<td>8</td>
<td>0.044318</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>s.e. of Coeff</th>
<th>t-ratio</th>
<th>prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-196.977</td>
<td>9.190</td>
<td>-21.4</td>
<td>≤ 0.0001</td>
</tr>
<tr>
<td>Year</td>
<td>0.100909</td>
<td>0.0046</td>
<td>21.8</td>
<td>≤ 0.0001</td>
</tr>
</tbody>
</table>

**Display 3.18** Data Desk output giving the equation of the least squares line for the minimum wage data.
You can ignore most of the output for now. You will learn how to interpret it in Chapter 11. For the time being, focus on the first two columns in the last three rows, which are reproduced in Display 3.19.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-196.977</td>
</tr>
<tr>
<td>Year</td>
<td>0.100909</td>
</tr>
</tbody>
</table>

Display 3.19 The lower-left corner of the computer output gives the y-intercept and slope.

The y-intercept is the coefficient in the row labeled “Constant” and is $-196.977$. The slope is the coefficient of the predictor variable “Year” and is $0.100909$. The SSE for the regression line is found in the “Residual” row and is 0.354545.

**DISCUSSION**

**Reading Computer Output**

D8. *Doctors’ incomes.* Display 3.20 shows the mean net income $y$ of family practitioners versus year $x$ (from page 121), with Data Desk computer output for the least squares line.

![Scatter Plot](image)

**Display 3.20** Scatterplot of mean net income (in thousands of dollars) of doctors board-certified in family practice, 1990–2001, and Data Desk output for the regression.
3.2 Getting a Line on the Pattern

a. What is the equation of the least squares line? Estimate the SSE from the scatterplot in Display 3.20, and then find it in the computer output.

b. The Minitab software output for this regression is shown in Display 3.21. How is it different from the Data Desk output?

Regression Analysis
The regression equation is
\[
\text{Income} = -8300.6 + 4.22 \text{Year}
\]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>Stdev</th>
<th>t-ratio</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-8300.6</td>
<td>933.4</td>
<td>8.89</td>
<td>0.000</td>
</tr>
<tr>
<td>Year</td>
<td>4.2248</td>
<td>0.4679</td>
<td>9.03</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\[ s = 4.774 \quad \text{R-sq} = 91.1\% \quad \text{R-sq(adj)} = 89.9\% \]

Analysis of Variance

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>1858.1</td>
<td>1858.1</td>
<td>81.52</td>
<td>0.000</td>
</tr>
<tr>
<td>Error</td>
<td>8</td>
<td>182.3</td>
<td>22.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>9</td>
<td>2040.4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Display 3.21 Minitab output for the regression of family practitioners’ income versus year.

Summary 3.2: Getting a Line on the Pattern

For many quantitative relationships, it makes sense to use one variable, \(x\), called the predictor or explanatory variable, to predict values of the other variable, \(y\), called the predicted or response variable. When the data are roughly linear, you can use a fitted line, called the least squares regression line, as a summary or model that describes the relationship between the two variables. You might also use it to predict the value of an unknown value \(y\) when you know the value of \(x\).

Interpolation—using a fitted relationship to predict a response value when the predictor value falls within the range of the data—generally is much more trustworthy than extrapolation—predicting response values based on the assumption that a fitted relationship applies outside the range of the observed data.

Each residual from a fitted line measures the vertical distance from a data point to the line:

\[ \text{residual} = \text{observed value} - \text{predicted value} = y - \hat{y} \]

The least squares regression line for a set of pairs \((x, y)\) is the line for which the sum of squared errors, or SSE, is as small as possible. For this line, these properties hold:

- The sum (and mean) of the residuals is 0.
- The line contains the point of averages, \((\bar{x}, \bar{y})\).
- The variation in the residuals is as small as possible.
- The line has slope \(b_1\), where
  \[
  b_1 = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}
  \]
To find the equation of the regression line,
- compute $\bar{x}$ and $\bar{y}$
- find the slope using the formula for $b_1$
- compute the $y$-intercept: $b_0 = \bar{y} - b_1\bar{x}$

The equation is $\hat{y} = b_0 + b_1x$. Remember to use a hat, $\hat{y}$, to indicate a predicted value of $y$.

### Practice

#### Lines as Summaries

P3. Display 3.22 shows the weight of a student’s pink eraser, in grams, plotted against the number of days into the school year. Estimate the slope of the line drawn on the graph. Interpret the slope in the context of the situation. [Source: Zach’s Eraser, CMC ComManiCator, 28 (June 2004): 28.]

Display 3.22  Weight of pink eraser.

P4. Display 3.23 shows the hand width of 383 students plotted against hand length. The line drawn on the plot is the least squares line.

Display 3.23 Hand width and hand length, in inches, for 383 students.

P5. If you attend a university where class sizes tend to be small, are you more likely to give to your alumni fund after you graduate than if you graduate from a university with large classes? Display 3.24 shows a scatterplot of a sample of 40 universities. Each university appears as a point. The vertical coordinate, $y$, tells the percentage of alumni who gave money. Each $x$-coordinate tells the student/faculty ratio (number of students per faculty member). The equation of the fitted line is approximately $\hat{y} = 55 - 2x$.

a. Which is the explanatory variable and which is the response variable?

b. Explain how you can see from the graph that an increase of five students per faculty member corresponds to a decrease of about 10 percentage points in the giving rate. Explain how you can see this from the equation of the fitted line.

c. Does the $y$-intercept have a useful interpretation in this situation?
d. Use the regression line to predict the giving rate for a university with a student/faculty ratio of 16. When you use the regression line to predict the giving rate, would you expect a rather large error or a relatively small error in your prediction?

e. Use the plot to estimate the residual for the university with the highest student/faculty ratio and for the university with the highest giving rate.

f. The university with the lowest student/faculty ratio, 6 to 1, had a giving rate of 32%. Use the equation of the fitted line to find the residual for that university.

g. Suppose the Alumni Association at Piranha State University boasts a giving rate of 80%. Without knowing the student/faculty ratio at PSU, can you tell whether the prediction error will be positive or negative?

c. Interpret the slope and y-intercept in the context of this situation.

d. Verify that the least squares regression line goes through the point of averages.

e. Verify that the sum of the residuals is 0.

P7. Use the statistical functions of your calculator to make a scatterplot, find the regression equation for predicting percentage on-time arrivals from mishandled baggage, and compute residuals for the airline data from P2 on page 110. [See Calculator Notes 3A, 3G, and 3D.] The data values are given in Display 3.25.

<table>
<thead>
<tr>
<th>Airline</th>
<th>Mishandled Baggage (per thousand passengers)</th>
<th>Percentage On-Time Arrivals</th>
</tr>
</thead>
<tbody>
<tr>
<td>America West</td>
<td>4.36</td>
<td>81.9</td>
</tr>
<tr>
<td>United</td>
<td>4.00</td>
<td>80.9</td>
</tr>
<tr>
<td>Southwest</td>
<td>4.42</td>
<td>78.4</td>
</tr>
<tr>
<td>US Airways</td>
<td>7.16</td>
<td>78.3</td>
</tr>
<tr>
<td>Continental</td>
<td>4.62</td>
<td>75.7</td>
</tr>
<tr>
<td>JetBlue</td>
<td>5.92</td>
<td>73.8</td>
</tr>
<tr>
<td>American</td>
<td>6.50</td>
<td>73.1</td>
</tr>
<tr>
<td>Delta</td>
<td>8.03</td>
<td>70.1</td>
</tr>
<tr>
<td>Alaska</td>
<td>7.02</td>
<td>69.1</td>
</tr>
<tr>
<td>Northwest</td>
<td>5.36</td>
<td>67.2</td>
</tr>
</tbody>
</table>

Display 3.25  Comparison, by airline, of mishandled baggage and on-time arrival rate.
[Source: U.S. Department of Transportation, Air Travel Consumer Report, October 2005.]

Display 3.24  Percentage of alumni giving to the alumni fund versus the student/faculty ratio for 40 highly rated U.S. universities.

Least Squares Regression Line

P6. The fat and calorie contents of 5 oz of three kinds of pizza are represented by the data points (9, 305), (11, 309), and (13, 316).

a. Plot the points.

b. Compute the equation of the least squares regression line by hand, and draw the line on your plot.
Reading Computer Output

P8. The JMP-IN computer output in Display 3.26 is for the pizza data in P6. Does it give the same results that you computed by hand? Where in the output is the SSE found?

![JMP-IN computer output for pizza data.](image)

Exercises

E9. Display 3.27 shows cost in dollars per hour versus number of seats for three aircraft models. Five lines, labeled A–E, are shown on the plot. Their equations, listed below, are labeled I–V.

a. Match each line (A–E) with its equation (I–V).
   I. \( \text{cost} = -290 + 15.8 \text{ seats} \)
   II. \( \text{cost} = 400 + 15.8 \text{ seats} \)
   III. \( \text{cost} = 1000 + 15.8 \text{ seats} \)
   IV. \( \text{cost} = -370 + 25 \text{ seats} \)
   V. \( \text{cost} = 900 + 10 \text{ seats} \)

b. Match each line (A–E) with the appropriate verbal description (I–V):
   I. This line overestimates cost.
   II. This line underestimates cost.
   III. This line overestimates cost for the smallest plane and underestimates cost for the largest plane.
   IV. This line underestimates cost for the smallest plane and overestimates cost for the largest plane.
   V. On balance, this line gives a better fit than the other lines.

![Cost in dollars per hour versus number of seats for three aircraft models.](image)
E10. Examine the scatterplot in Display 3.28.

**Display 3.28** *Calories versus fat, per 5-oz serving, for seven kinds of pizza.* [Source: Consumer Reports, July 2003.]

a. Which two kinds of pizza in Display 3.28 have the fewest calories? Which two have the least fat? Which region of the graph has the pizzas with the most fat?

b. Display 3.29 shows the data again, with five possible summary lines. Match each equation (I–V) with the appropriate line (A–E).

I. \(\text{calories} = 70 + 15 \text{ fat}\)
II. \(\text{calories} = -10 + 25 \text{ fat}\)
III. \(\text{calories} = 150 + 15 \text{ fat}\)
IV. \(\text{calories} = 110 + 15 \text{ fat}\)
V. \(\text{calories} = 170 + 10 \text{ fat}\)

c. Consider the possible summary lines in Display 3.29.
   i. Which line gives predicted values for calorie content that are too high? How can you tell this from the plot?
   ii. Which line tends to give predicted calorie values that are too low?
   iii. Which line tends to overestimate calorie content for lower-fat pizzas and underestimate calorie content for higher-fat pizzas?
   iv. Which line has the opposite problem, underestimating calorie content when fat content is lower and overestimating calorie content when fat content is higher?
   v. Which line fits the data best overall?

E11. *Heights of boys.* The scatterplot in Display 3.30 shows the median height, in inches, for boys ages 2 through 14 years.


a. Estimate the slope of the line that summarizes the relationship between age and median height.

b. Explain the meaning of the slope with respect to boys and their median height.

c. Write the equation of the line using the slope from part a and a point on the line.

d. Interpret the y-intercept. Does the interpretation make sense in this context?
E12. *Pizza again.* Display 3.31 shows the calorie and fat content of 5 oz of various kinds of pizza.

<table>
<thead>
<tr>
<th>Pizza</th>
<th>Fat (g)</th>
<th>Calories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pizza Hut’s Hand Tossed</td>
<td>9</td>
<td>230</td>
</tr>
<tr>
<td>Domino’s Deep Dish</td>
<td>19.5</td>
<td>385</td>
</tr>
<tr>
<td>Pizza Hut’s Pan</td>
<td>14</td>
<td>280</td>
</tr>
<tr>
<td>Domino’s Hand Tossed</td>
<td>12</td>
<td>305</td>
</tr>
<tr>
<td>Little Caesar’s Original Round</td>
<td>8</td>
<td>230</td>
</tr>
<tr>
<td>Little Caesar’s Deep Dish</td>
<td>14.2</td>
<td>350</td>
</tr>
<tr>
<td>Pizza Hut’s Stuffed Crust</td>
<td>15</td>
<td>370</td>
</tr>
</tbody>
</table>

Display 3.31 Calories and fat content per 5-oz serving, for seven kinds of pizza. [Source: Consumer Reports, January 2002.]

a. Use the line on the scatterplot to predict the calorie content of a pizza with 10.5 g of fat. Then use the line to predict the calorie content of a pizza with 15 g of fat.

b. Use the two predictions in part a to estimate the slope of the line. Write the equation of the line using this slope and a point on the line.

c. There are 9 calories in a gram of fat. How is your estimated slope related to this number?

E13. *Stopping on a dime?* In an emergency, the typical driver requires about 0.75 second to get his or her foot onto the brake pedal. The distance the car travels during this reaction time is called the *reaction distance*. Display 3.32 shows the reaction distances for cars traveling at various speeds.

<table>
<thead>
<tr>
<th>Speed (mi/h)</th>
<th>Reaction Distance (ft)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>22</td>
</tr>
<tr>
<td>30</td>
<td>33</td>
</tr>
<tr>
<td>40</td>
<td>44</td>
</tr>
<tr>
<td>50</td>
<td>55</td>
</tr>
<tr>
<td>60</td>
<td>66</td>
</tr>
<tr>
<td>70</td>
<td>77</td>
</tr>
</tbody>
</table>

Display 3.32 Reaction distance at various speeds.

a. Plot *reaction distance* versus *speed*, with *speed* on the horizontal axis. Describe the shape of the plot.

b. What should the *y*-intercept be?

c. Find the slope of the line of best fit by calculating the change in *y* per unit change in *x*. What does the slope represent in this situation?

d. Write the equation of the line that fits these data.

e. Use the equation of the line in part d to predict the reaction distance for a car traveling at a speed of 55 mi/h and at 75 mi/h.

f. How would the equation change if it actually took 1 second, instead of 0.75 second, for drivers to react?

E14. The scatterplot in Display 3.33 shows *operating cost* (in dollars per hour) versus *fuel consumption* (in gallons per hour) for a sample of commercial aircraft.

Display 3.33 Operating cost versus fuel consumption for commercial aircraft.
a. Which is the explanatory variable and which is the response variable?
b. Estimate the slope of the regression line from the graph, and interpret it in the context of this situation.
c. The y-intercept is 470. Does this value have a reasonable interpretation in this situation?
d. Use the line to predict the cost per hour for a plane that consumes 1500 gal/h of fuel.

E15. Arsenic is a potent poison sometimes found in groundwater. Long-term exposure to arsenic in drinking water can cause cancer. How much arsenic a person has absorbed can be measured from a toenail clipping. The plot in Display 3.34 shows the arsenic concentrations in the toenails of 21 people who used water from their private wells plotted against the arsenic concentration in their well water. Both measurements are in parts per million.


a. What is the predictor variable, and what is the response variable?
b. Describe the relationship.
c. Estimate the residual for the person with the highest concentration of arsenic in the well water.
d. Find the person on the plot with the largest residual. What was the concentration of arsenic in that person’s toenails?
e. The World Health Organization has set a standard that the concentration of arsenic in drinking water should be less than 0.01 mg/L. (1 mg/L = 1 ppm.) Is this standard exceeded in any of these wells? [Source: www.who.int.]

a. The least squares residuals for the pizza data are, in order from smallest to largest, \(-40.58, -17.66, -15.95, -1.03, 14.28, 26.44, \) and \(34.50.\) Match each residual with its pizza.
b. What does the residual for Pizza Hut’s Pan pizza tell you about the pizza’s number of calories versus fat content?
c. For Pizza Hut’s Hand Tossed and Domino’s Deep Dish, are the residuals positive or negative? How can you tell this from the scatterplot in Display 3.31?

E17. The level of air pollution is indicated by a measure called the air quality index (AQI). An AQI greater than 100 means the air quality is unhealthy for sensitive groups such as children. The table and plot in Display 3.35 show the number of days in Detroit that the AQI was greater than 100 for the years 2001, 2002, and 2003.


a. By hand, compute the equation of the least squares line.
b. Interpret the slope in the context of this situation.
c. Which year has the largest residual? What is this residual?
d. Compute the SSE for this line.
e. Verify that the sum of the residuals is 0.
f. Find the SSE for the line that has the same slope as the least squares line but passes through the point for 2002. Is this SSE larger or smaller than the SSE for the least squares line? According to the least squares approach, which line fits better?
g. Find the slope of the line that passes through the points for 2001 and 2003. Then find the fitted value for 2002. Finally, find all three residuals and the value of the SSE for this line.
h. The least squares line doesn't pass through any of the points, and yet judging by the SSE that line fits better than the one in part g. Do you agree that the least squares line fits better than the lines in parts f and g? Explain why or why not.

E18. Even more pizza. Refer again to the table and scatterplot in Display 3.31 on page 134.

a. By hand, compute the equation of the least squares regression line for using fat to predict calories. How close was your estimate of the equation in E12?
b. Which of these values must be the SSE for this regression? Explain your answer.


a. Practice using your calculator by making a scatterplot, finding the equation of the least squares line for median height versus age, and graphing the equation on the plot.

<table>
<thead>
<tr>
<th>Age (yr)</th>
<th>Median Height (in.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>35.1</td>
</tr>
<tr>
<td>3</td>
<td>38.7</td>
</tr>
<tr>
<td>4</td>
<td>41.3</td>
</tr>
<tr>
<td>5</td>
<td>44.1</td>
</tr>
<tr>
<td>6</td>
<td>46.5</td>
</tr>
<tr>
<td>7</td>
<td>48.6</td>
</tr>
<tr>
<td>8</td>
<td>51.7</td>
</tr>
<tr>
<td>9</td>
<td>53.7</td>
</tr>
<tr>
<td>10</td>
<td>56.1</td>
</tr>
<tr>
<td>11</td>
<td>59.5</td>
</tr>
<tr>
<td>12</td>
<td>61.2</td>
</tr>
<tr>
<td>13</td>
<td>62.9</td>
</tr>
<tr>
<td>14</td>
<td>63.6</td>
</tr>
</tbody>
</table>

Display 3.36 Median height for girls ages 2–14.

b. Judging from the plot, is the residual for 11-year-olds positive or negative? Compute this residual to check your answer.
c. Verify that the line contains the point of averages, \((\bar{x}, \bar{y})\).
d. How does the regression line for girls compare to the line for boys in E11?

E20. Sum of residuals. In this exercise, you will show that the sum of the residuals is equal to 0 if and only if the regression line passes through the point of averages, \((\bar{x}, \bar{y})\).

a. Show that for a horizontal line the sum of the residuals will be 0 if and only if the line passes through the point of averages.
b. Show that no matter what the slope of the line is, the sum of the residuals will be 0 if and only if the line passes through the point of averages.
c. Why isn't it good enough to define the regression line as the line that makes the sum of the residuals equal 0?

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>Stdev</th>
<th>t-ratio</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>31.59</td>
<td>0.31</td>
<td>103.17</td>
<td>0.00</td>
</tr>
<tr>
<td>Age</td>
<td>2.47</td>
<td>0.03</td>
<td>71.34</td>
<td>0.00</td>
</tr>
</tbody>
</table>

\[ s = 0.47 \quad R^2 = 99.8\% \quad R^2(adj) = 99.8\% \]

Analysis of Variance

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>1114.1</td>
<td>1114.1</td>
<td>5089.73</td>
<td>0.000</td>
</tr>
<tr>
<td>Error</td>
<td>11</td>
<td>2.4</td>
<td>0.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>12</td>
<td>1116.5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Display 3.37** Computer output of median height versus age data.

a. Write the equation of the regression line. How does it compare to your estimate of the equation in E11?
b. What is the SSE for this least squares line? Does its value seem reasonable given the scatterplot in Display 3.30 on page 133?

e22. Part of a printout for the percentage of alumni who give to their colleges versus the student/faculty ratio is shown in Display 3.38. (These are the data in the scatterplot shown in Display 3.24 on page 131.)

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>Stdev</th>
<th>t-ratio</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>54.98</td>
<td>5.48</td>
<td>10.04</td>
<td>0.00</td>
</tr>
<tr>
<td>S/F Rati</td>
<td>-1.95</td>
<td>0.44</td>
<td>-4.47</td>
<td>0.00</td>
</tr>
</tbody>
</table>

\[ s = 9.70 \quad R^2 = 34.4\% \quad R^2(adj) = 32.7\% \]

Analysis of Variance

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>1880.2</td>
<td>1880.2</td>
<td>19.97</td>
<td>0.000</td>
</tr>
<tr>
<td>Error</td>
<td>38</td>
<td>3578.5</td>
<td>94.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>39</td>
<td>5458.7</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Unusual Observations

<table>
<thead>
<tr>
<th>Obs</th>
<th>S/F Rati</th>
<th>Alumni</th>
<th>Fit</th>
<th>Fit Residual</th>
<th>Resid</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>10.0</td>
<td>14.0</td>
<td>35.52</td>
<td>1.78</td>
<td>-2.26</td>
</tr>
<tr>
<td>14</td>
<td>11.0</td>
<td>54.0</td>
<td>33.58</td>
<td>1.60</td>
<td>20.42</td>
</tr>
<tr>
<td>30</td>
<td>13.0</td>
<td>52.0</td>
<td>29.69</td>
<td>1.59</td>
<td>22.31</td>
</tr>
<tr>
<td>39</td>
<td>20.0</td>
<td>15.0</td>
<td>16.07</td>
<td>3.78</td>
<td>-1.07</td>
</tr>
<tr>
<td>40</td>
<td>24.0</td>
<td>9.0</td>
<td>8.29</td>
<td>5.41</td>
<td>0.71</td>
</tr>
</tbody>
</table>

**Display 3.38** Computer output: regression analysis of percentage giving to alumni fund versus student/faculty ratio.

a. What equation is given in the printout for the least squares regression line?
b. Examine the table of unusual observations. What is the student/faculty ratio at the college with the largest residual (in absolute value)? Find this college in Display 3.24 on page 131.
c. Verify that the fit and the value of the largest residual were computed correctly.
d. Locate the SSE on the printout. Why is this value so large?

e23. For the least squares regression line you found in E19, calculate the residuals for girls ages 2, 8, and 14. What does this suggest about the pattern of growth beyond what is summarized in the equation of the regression line?

e24. More about slope.

a. You and three friends, one right after the other, each buy the same kind of gas at the same pump. Then you make a scatterplot of your data, with one point per person, plotting the number of gallons on the x-axis and the total price paid on the y-axis. Will all four points lie on the same line? Explain.
b. You and the same three friends each drive 80 mi but at different average speeds. Afterward, you plot your data twice, first as a set of four points with coordinates average speed, x, and elapsed time, y, and then as a set of points with coordinates average speed, x, and y* defined as \( \frac{1}{\text{elapsed time}} \). Which plot will give a straight line? Explain your reasoning. Will the other plot be a curve opening up, a curve opening down, or neither?

e25. The data set in Display 3.39 is the pizza data of E12 augmented by other brands of cheese pizza typically sold in supermarkets.

a. Plot calories versus fat. Does there appear to be a linear association between calories and fat? If so, fit a least squares line to the data, and interpret the slope of the line.
b. Plot fat versus cost. Does there appear to be a linear association between cost and fat? If so, fit a least squares line to the data, and interpret the slope of the line.

c. Plot calories versus cost. Does there appear to be a linear association between cost and calories?

d. Write a summary of your findings.

Display 3.39  Food values and cost per 5-oz serving of pizza. [Source: Consumer Reports, January 2002.]

<table>
<thead>
<tr>
<th>Pizza</th>
<th>Calories</th>
<th>Fat (g)</th>
<th>Cost ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domino's Deep Dish</td>
<td>385</td>
<td>19.5</td>
<td>1.87</td>
</tr>
<tr>
<td>Pizza Hut's Stuffed Crust</td>
<td>370</td>
<td>15</td>
<td>1.83</td>
</tr>
<tr>
<td>Pizza Hut's Pan</td>
<td>280</td>
<td>14</td>
<td>1.83</td>
</tr>
<tr>
<td>Domino's Hand Tossed</td>
<td>305</td>
<td>12</td>
<td>1.67</td>
</tr>
<tr>
<td>Pizza Hut's Hand Tossed</td>
<td>230</td>
<td>9</td>
<td>1.63</td>
</tr>
<tr>
<td>Little Caesar’s Deep Dish</td>
<td>350</td>
<td>14.2</td>
<td>1.06</td>
</tr>
<tr>
<td>Little Caesar’s Original Round</td>
<td>230</td>
<td>8</td>
<td>0.81</td>
</tr>
<tr>
<td>Freschetta Bakes &amp; Rises 4-Cheese</td>
<td>364</td>
<td>15</td>
<td>0.98</td>
</tr>
<tr>
<td>Freschetta Bakes &amp; Rises Sauce Stuffed Crust 4-Cheese</td>
<td>334</td>
<td>11</td>
<td>1.23</td>
</tr>
<tr>
<td>DiGiorno Rising Crust Four Cheese</td>
<td>332</td>
<td>12</td>
<td>0.94</td>
</tr>
<tr>
<td>Amy’s Organic Crust &amp; Tomatoes Cheese</td>
<td>341</td>
<td>14</td>
<td>1.92</td>
</tr>
<tr>
<td>Safeway Select Verdi Quattro Fromaggio Self Rising Crust</td>
<td>307</td>
<td>9</td>
<td>0.84</td>
</tr>
<tr>
<td>Tony’s Super Rise Crust Four Cheese</td>
<td>335</td>
<td>12</td>
<td>0.96</td>
</tr>
<tr>
<td>Kroger Self Rising Crust Four Cheese</td>
<td>292</td>
<td>9</td>
<td>0.80</td>
</tr>
<tr>
<td>Tombstone Stuffed Crust Cheese</td>
<td>364</td>
<td>18</td>
<td>0.96</td>
</tr>
<tr>
<td>Red Baron Classic 4 Cheese</td>
<td>384</td>
<td>20</td>
<td>0.91</td>
</tr>
<tr>
<td>Boboli Original</td>
<td>333</td>
<td>12</td>
<td>0.89</td>
</tr>
<tr>
<td>Tombstone Original Extra Cheese</td>
<td>328</td>
<td>14</td>
<td>0.94</td>
</tr>
<tr>
<td>Reggio’s Chicago Style Cheese</td>
<td>367</td>
<td>13</td>
<td>1.02</td>
</tr>
<tr>
<td>Jack’s Original Cheese</td>
<td>325</td>
<td>13</td>
<td>0.92</td>
</tr>
<tr>
<td>Celeste Pizza for One Cheese</td>
<td>346</td>
<td>17</td>
<td>1.17</td>
</tr>
<tr>
<td>McCain Elliot’s Cheese</td>
<td>299</td>
<td>9</td>
<td>0.54</td>
</tr>
<tr>
<td>Michelina’s Zap ’ems That’za Pizza!</td>
<td>394</td>
<td>19</td>
<td>1.28</td>
</tr>
<tr>
<td>Totino’s The Original Crisp Crust Party Cheese</td>
<td>322</td>
<td>14</td>
<td>0.67</td>
</tr>
</tbody>
</table>

E26. Poverty. What variables are most closely associated with poverty? Display 3.40 provides information on population characteristics of the 50 U.S. states plus the District of Columbia. Each variable is measured as a percentage of the state’s population, as described here:

- Percentage living in metropolitan areas
- Percentage white
- Percentage of adults who have graduated from high school
- Percentage of families with incomes below the poverty line
- Percentage of families headed by a single parent

Construct scatterplots to determine which variables are most strongly associated with poverty.

Write a letter to your representative in Congress about poverty in America, relying only on what you find in these data. Point out the variables that appear to be most strongly associated with poverty and those that appear to have little or no association with poverty.
<table>
<thead>
<tr>
<th>State</th>
<th>Metropolitan Residence</th>
<th>White Graduates</th>
<th>Poverty</th>
<th>Single Parent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alabama</td>
<td>55.4</td>
<td>71.3</td>
<td>79.9</td>
<td>14.6</td>
</tr>
<tr>
<td>Alaska</td>
<td>65.6</td>
<td>70.8</td>
<td>90.6</td>
<td>8.3</td>
</tr>
<tr>
<td>Arizona</td>
<td>88.2</td>
<td>87.7</td>
<td>83.8</td>
<td>13.3</td>
</tr>
<tr>
<td>Arkansas</td>
<td>52.5</td>
<td>81</td>
<td>80.9</td>
<td>18</td>
</tr>
<tr>
<td>California</td>
<td>94.4</td>
<td>77.5</td>
<td>81.1</td>
<td>12.8</td>
</tr>
<tr>
<td>Colorado</td>
<td>84.5</td>
<td>90.2</td>
<td>88.7</td>
<td>9.4</td>
</tr>
<tr>
<td>Connecticut</td>
<td>87.7</td>
<td>85.4</td>
<td>87.5</td>
<td>7.8</td>
</tr>
<tr>
<td>Delaware</td>
<td>80.1</td>
<td>76.3</td>
<td>88.7</td>
<td>8.1</td>
</tr>
<tr>
<td>District of Columbia</td>
<td>100</td>
<td>36.2</td>
<td>86</td>
<td>16.8</td>
</tr>
<tr>
<td>Florida</td>
<td>89.3</td>
<td>80.6</td>
<td>84.7</td>
<td>12.1</td>
</tr>
<tr>
<td>Alabama</td>
<td>55.4</td>
<td>71.3</td>
<td>79.9</td>
<td>14.6</td>
</tr>
<tr>
<td>Alaska</td>
<td>65.6</td>
<td>70.8</td>
<td>90.6</td>
<td>8.3</td>
</tr>
<tr>
<td>Arizona</td>
<td>88.2</td>
<td>87.7</td>
<td>83.8</td>
<td>13.3</td>
</tr>
<tr>
<td>Arkansas</td>
<td>52.5</td>
<td>81</td>
<td>80.9</td>
<td>18</td>
</tr>
<tr>
<td>California</td>
<td>94.4</td>
<td>77.5</td>
<td>81.1</td>
<td>12.8</td>
</tr>
<tr>
<td>Colorado</td>
<td>84.5</td>
<td>90.2</td>
<td>88.7</td>
<td>9.4</td>
</tr>
<tr>
<td>Connecticut</td>
<td>87.7</td>
<td>85.4</td>
<td>87.5</td>
<td>7.8</td>
</tr>
<tr>
<td>Delaware</td>
<td>80.1</td>
<td>76.3</td>
<td>88.7</td>
<td>8.1</td>
</tr>
<tr>
<td>District of Columbia</td>
<td>100</td>
<td>36.2</td>
<td>86</td>
<td>16.8</td>
</tr>
<tr>
<td>Florida</td>
<td>89.3</td>
<td>80.6</td>
<td>84.7</td>
<td>12.1</td>
</tr>
</tbody>
</table>

Display 3.40  Characteristics of state populations, as percentage of population.

[Source: U.S. Census Bureau, www.census.gov.]
Some of the linear relationships you’ve seen in this chapter have been extremely strong, with points packed tightly around the regression line. Other linear relationships have been quite weak—although there was a general linear trend, a lot of variation was present in the values of $y$ associated with a given value of $x$. Still other linear relationships have been in between.

In this section, you’ll learn how to measure the strength of a linear relationship numerically by using the correlation coefficient, $r$ (which, from this point on, will be referred to simply as the correlation), where $-1 \leq r \leq 1$. Just as in the last section, you’ll start by working intuitively and visually and then move on to a computational approach.

### Estimating the Correlation

Examine the scatterplots and their correlations in Display 3.41. To get a rough idea of the size of a correlation, it is helpful to sketch an ellipse around the cloud of points in the scatterplot. If the ellipse has points scattered throughout and the points appear to follow a linear trend, then the correlation is a reasonable measure of the strength of the association. If the ellipse slants upward as you go from left to right, the correlation is positive. If the ellipse slants downward as you go from left to right, the correlation is negative. If the ellipse is fat, the correlation is weak and the absolute value of $r$ is close to 0. If the ellipse is skinny, the correlation is strong and the absolute value of $r$ is close to 1.

Display 3.41  Scatterplots with ellipses and their correlations.

In Activity 3.3a, you will learn more about correlation and you’ll practice finding the value of $r$ using your calculator.
ACTIVITY 3.3a

Was Leonardo Correct?

What you’ll need: a measuring tape, yardstick, or meterstick

Leonardo da Vinci was a scientist and an artist who combined these skills to draft extensive instructions for other artists on how to proportion the human body in painting and sculpture. Three of Leonardo’s rules were:
- height is equal to the span of the outstretched arms
- kneeling height is three-fourths of the standing height
- the length of the hand is one-ninth of the height

1. Work with a partner to measure your height, kneeling height, arm span, and hand length. Combine your data with the rest of your class.
2. Check Leonardo’s three rules visually by plotting the data on three scatterplots.
3. For the plots that have a linear trend, use your calculator to find the equation of the regression line and the value of \( r \) (the correlation). [See Calculator Notes 3G and 3H.]
4. Interpret the slopes of the regression lines. Interpret the correlations.
5. Do the three relationships described by Leonardo appear to hold? Do they hold strongly?

DISCUSSION

Estimating the Correlation

D9. Match each of the four scatterplots with its correlation, choosing from the values –0.783, 0.783, 0.908, and 0.999.
   a. year of birth versus year of hire (Display 3.1 on page 106)
   b. age at layoffs versus year of hire (Display 3.2 on page 106)
   c. median height of boys versus age (Display 3.30 on page 133)
   d. calories in pizza versus fat (Display 3.31 on page 134)

D10. Four relationships are described here.
   I. For a random sample of students from the senior class, \( x \) represents the day of the month of the person’s birthday and \( y \) represents the cost of the person’s most recent haircut.
   II. For a random collection of U.S. coins, \( x \) represents the diameter and \( y \) represents the circumference.
   III. For a random sample of bags of white socks, \( x \) represents the number of socks and \( y \) represents the price per bag.
   IV. For a random sample of bags of white socks, \( x \) represents the number of socks and \( y \) represents the price per sock.
   a. Which of these relationships have a positive correlation, and which a negative correlation? Which has the strongest relationship? The weakest?
   b. For each of the four relationships, discuss the connection between your ability to precisely predict a value of \( y \) for a given value of \( x \) and the strength of the correlation.
A Formula for the Correlation, $r$

A formula for the correlation, $r$, follows. It looks impressive, but the basic idea is simple—you convert $x$ and $y$ to standardized values ($z$-scores), and then find their average product (dividing by $n - 1$).

$$ r = \frac{1}{n-1} \sum \left( \frac{x - \bar{x}}{s_x} \right) \left( \frac{y - \bar{y}}{s_y} \right) = \frac{1}{n-1} \sum z_x \cdot z_y $$

In this formula, $s_x$ is the standard deviation of the $x$'s, and $s_y$ is the standard deviation of the $y$'s. Remember that the $z$-score tells you how many standard deviations the value lies above or below the mean.

Example: Computing $r$ for the Airline Data

Compute the correlation for the relationship between the number of mishandled bags per thousand passengers and the percentage of on-time arrivals for the airline data in Display 3.25 on page 131.

Solution

When you are on a desert island or taking a test where you must compute $r$ by hand, it is easiest to organize your work as in Display 3.42. First, compute the average values of $x$ and $y$, $\bar{x}$ and $\bar{y}$. Then compute their standard deviations, $s_x$ and $s_y$.

$$ \bar{x} = \frac{57.39}{10} = 5.739 \quad \bar{y} = \frac{748.5}{10} = 74.85 $$

$$ s_x = 1.3977 \quad s_y = 5.0535 $$

<table>
<thead>
<tr>
<th>Airline</th>
<th>Mishandled Baggage (per thousand passengers)</th>
<th>Percentage On-Time Arrivals</th>
<th>$z_x = \frac{x - \bar{x}}{s_x}$</th>
<th>$z_y = \frac{y - \bar{y}}{s_y}$</th>
<th>$z_x \cdot z_y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>America West</td>
<td>4.36</td>
<td>81.9</td>
<td>$-0.98662$</td>
<td>$1.39506$</td>
<td>$-1.37640$</td>
</tr>
<tr>
<td>United</td>
<td>4.00</td>
<td>80.9</td>
<td>$-1.24419$</td>
<td>$1.19718$</td>
<td>$-1.48951$</td>
</tr>
<tr>
<td>Southwest</td>
<td>4.42</td>
<td>78.4</td>
<td>$-0.94369$</td>
<td>$0.70248$</td>
<td>$-0.66292$</td>
</tr>
<tr>
<td>US Airways</td>
<td>7.16</td>
<td>78.3</td>
<td>$1.01667$</td>
<td>$0.68269$</td>
<td>$0.69407$</td>
</tr>
<tr>
<td>Continental</td>
<td>4.62</td>
<td>75.7</td>
<td>$-0.80060$</td>
<td>$0.16820$</td>
<td>$-0.13466$</td>
</tr>
<tr>
<td>JetBlue</td>
<td>5.92</td>
<td>73.8</td>
<td>$0.12950$</td>
<td>$-0.20777$</td>
<td>$-0.02691$</td>
</tr>
<tr>
<td>American</td>
<td>6.50</td>
<td>73.1</td>
<td>$0.54447$</td>
<td>$-0.34629$</td>
<td>$-0.18854$</td>
</tr>
<tr>
<td>Delta</td>
<td>8.03</td>
<td>70.1</td>
<td>$1.63912$</td>
<td>$-0.93993$</td>
<td>$-1.54067$</td>
</tr>
<tr>
<td>Alaska</td>
<td>7.02</td>
<td>69.1</td>
<td>$0.91651$</td>
<td>$-1.13781$</td>
<td>$-1.04281$</td>
</tr>
<tr>
<td>Northwest</td>
<td>5.36</td>
<td>67.2</td>
<td>$-0.27116$</td>
<td>$-1.51379$</td>
<td>$0.41048$</td>
</tr>
<tr>
<td>Sum</td>
<td>57.39</td>
<td>748.5</td>
<td>0</td>
<td>0</td>
<td>$-5.35787$</td>
</tr>
<tr>
<td>Mean</td>
<td>5.739</td>
<td>74.85</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>1.3977</td>
<td>5.0535</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Display 3.42 Calculations for the correlation for airline data.
Then convert each value of $x$ and $y$ to standard units, or $z$-scores. The correlation, $r$, is the average product of the $z$-scores—the sum of the products in the last column divided by $n - 1$.

\[
r = \frac{1}{n - 1} \sum z_x \cdot z_y
\]

\[
= \frac{1}{10 - 1} (-5.35787)
\]

\[
= -0.595
\]

[You can use a calculator to find the value of $r$. See Calculator Note 3H.]

A way to visualize the computations in the example is to look at the four quadrants formed on the scatterplot by dividing it vertically at the mean value of $x$ and horizontally at the mean value of $y$. Such a plot is shown in Display 3.43. Points in Quadrant I, such as the point for US Airways, have positive $z$-scores for both $x$ and $y$, so their product contributes a positive amount to the calculation of the correlation. Points in Quadrant III, such as the point for Northwest, have negative $z$-scores for both $x$ and $y$, so their product also contributes a positive amount to the calculation of the correlation. Points in Quadrants II and IV, such as America West and Delta, contribute negative amounts to the calculation of the correlation because one $z$-score is positive and the other is negative.

Display 3.43 Scatterplot divided into quadrants at $(x, y)$.

Because $r$ is the average of the products $z_x \cdot z_y$ and $z$-scores have no units, $r$ has no units. In fact, $r$ does not depend on the units of measurement in the original data. In Activity 3.3a, if you measure arm spans and heights in inches and your friend measures them in centimeters, you will both compute the same value for the correlation.
A Formula for the Correlation, \( r \)

D11. Look at Display 3.42, showing the calculations for the correlation, \( r \).
   a. Confirm the calculations for the first row, America West.
   b. Which point makes the largest contribution to the correlation? Where is this point on the scatterplot?
   c. Which point makes the smallest contribution to the correlation? Where is this point on the scatterplot?

D12. Understanding \( r \).
   a. Explain in your own words what the correlation measures.
   b. Explain in your own words why \( r \) has no units.
   c. When computing the correlation between two variables, does it matter which variable you select as \( y \) and which you select as \( x \)? Explain.

D13. Refer to Display 3.43.
   a. What can you say about \( r \) if there are many points in Quadrants I and III and few in Quadrants II and IV?
   b. What can you say about \( r \) if there are many points in Quadrants II and IV and few in Quadrants I and III?
   c. What can you say about \( r \) if points are scattered randomly in all four quadrants?

Correlation and the Appropriateness of a Linear Model

It is tempting to believe that a high correlation (either positive or negative) is evidence that a linear model is appropriate for your data. Alas, the real world is not so simple. For example, Display 3.44 shows the number of blogs (Web-based periodic postings of a person’s thoughts) for the first few years after 2003, when blogging’s popularity took off. The growth is exponential, yet the (linear) correlation is very strong, \( r = 0.91 \). The points do cluster fairly tightly about the linear regression line, but that line is not the best model for the data. The plot on the right shows the points and the graph of the best-fitting exponential equation, \( y = 0.353 \cdot 1.140^x \). (You will learn how to fit exponential equations to data in Section 3.5.)

<table>
<thead>
<tr>
<th>Month</th>
<th>Months from March 2003</th>
<th>Number of Blogs (millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aug. 2003</td>
<td>5</td>
<td>0.6</td>
</tr>
<tr>
<td>Jan. 2004</td>
<td>10</td>
<td>1.5</td>
</tr>
<tr>
<td>June 2004</td>
<td>15</td>
<td>2.7</td>
</tr>
<tr>
<td>Nov. 2004</td>
<td>20</td>
<td>4.6</td>
</tr>
<tr>
<td>Apr. 2005</td>
<td>25</td>
<td>9.3</td>
</tr>
<tr>
<td>Sept. 2005</td>
<td>30</td>
<td>17.8</td>
</tr>
</tbody>
</table>
3.3 Correlation: The Strength of a Linear Trend

Blog growth with a linear regression. Blog growth with graph of an exponential equation, $y = 0.353 \cdot 1.140^x$.


The quiz scores for 22 students in Display 3.45, on the other hand, have a correlation, $r$, of only 0.48. There is quite a bit of scatter, partly because the quizzes covered very different topics. Quiz 2 covered exponential growth, and Quiz 3 covered probability. In spite of the scatter, a line is the most appropriate model because there is no curvature in the pattern of data points.

**Display 3.45** Scores on two consecutive 30-point quizzes.

---

**DISCUSSION**

**Correlation and the Appropriateness of a Linear Model**

D14. When the correlation is small in absolute value, what does it mean for the prediction error? Why would anyone want to fit a line to data in a case in which the correlation is small, as in Display 3.45 (quiz scores)?

D15. Provide a real-life scenario involving two variables for each situation. Assume $r$ is positive in each case.

a. $r$ is small and you do not want to fit a line.
b. $r$ is small and you do want to fit a line.
c. $r$ is large and you do not want to fit a line.
d. $r$ is large and you do want to fit a line.
D16. It’s common in situations similar to the growth in blogs for the numbers to increase exponentially for the first few years. What do you think would happen if the table in Display 3.44 were continued to include numbers for months up to the current year?

The Relationship Between the Correlation and the Slope

By now you might have observed that the slope of the regression line, \( b_1 \), and the correlation, \( r \), always have the same sign. But these have a more specific relationship.

Finding the Slope from the Correlation and the SDs

The slope of a least squares regression line, \( b_1 \), and the correlation, \( r \), are related by the equation

\[
b_1 = r \frac{s_y}{s_x}
\]

where \( s_x \) is the standard deviation of the \( x \)'s and \( s_y \) is the standard deviation of the \( y \)'s. This means that if you standardize the data so that \( s_x = 1 \) and \( s_y = 1 \), then the slope of the regression line is equal to the correlation.

Example: Critical Reading and Math SAT Scores

In 2005, the mean critical reading score for all SAT I test takers was 508, with a standard deviation of 113. For math scores, the mean was 520, with a standard deviation of 115. The correlation between the two scores was not given but is known to be quite high. If you can estimate this correlation as, say, 0.7, you can find the equation of the regression line and use it to estimate the math score from a student’s critical reading score. [Source: The College Board, 2005 College Bound Seniors: A Profile of SAT Program Test Takers.]

Solution

The formula gives an estimate of the slope of

\[
b_1 = r \frac{s_y}{s_x} = 0.7 \cdot \frac{115}{113} \approx 0.71
\]

To find the \( y \)-intercept, use the fact that the point \((\bar{x}, \bar{y}) = (508, 520)\) is on the regression line:

\[
y = \text{slope} \cdot x + \text{\( y \)-intercept}
\]

\[
520 = 0.71 \cdot (508) + \text{\( y \)-intercept}
\]

\[
\text{\( y \)-intercept} = 159.32
\]

The equation is \( \hat{y} = 0.71x + 159.32 \).
Correlation: The Strength of a Linear Trend

The Relationship Between the Correlation and the Slope

D17. Find the equation of the regression line for predicting an SAT I critical reading score given the student’s SAT I math score.

Correlation Does Not Imply Causation

In a sample of elementary school students, there is a strong positive relationship between shoe size and scores on a standardized test of ability to do arithmetic. Does this mean that studying arithmetic makes your feet bigger? No. Shoe size and skill at arithmetic are related to each other because both increase as a child gets older. Age is an example of a lurking variable.

Beware the Lurking Variable

A lurking variable is a variable that you didn’t include in your analysis but that might explain the relationship between the variables you did include. That is, when variables $x$ and $y$ are correlated, it might be because both are consequences of a third variable, $z$, that is lurking in the background.

Even if you can’t identify a lurking variable, you should be careful to avoid jumping to a conclusion about cause and effect when you observe a strong relationship. The value of $r$ does not tell you anything about why two variables are related. The statement “Correlation does not imply causation” can help you remember this. To conclude that one thing causes another, you need data from a randomized experiment, as you’ll learn in the next chapter.

Correlation Does Not Imply Causation

D18. Display 3.43 on page 143 shows a negative association between the percentage of on-time arrivals and the number of mishandled bags per thousand passengers. Discuss whether you think one of these variables might cause the other, or whether a lurking variable might account for both.

D19. For the sample of 50 top-rated universities in E5 on page 113, there’s a very strong positive relationship between acceptance rate (percentage of applicants who are offered admission) and SAT scores (the 75th percentile for an entering class). Explain why these two variables have such a strong relationship. Does one “cause” the other? If not, how might you account for the strong relationship?

D20. People who argue about politics and public policy often point to relationships between quantitative variables and then offer a cause-and-effect explanation to support their points of view. For each of these relationships, first give a possible explanation by assuming a causal relationship and then give another possible explanation based on a lurking variable.

a. Faculty positions in academic subjects with a higher percentage of male faculty tend to pay higher salaries. (For example, engineering and geology have high percentages of male faculty and high average salaries;
journalism, music, and social work have much lower percentages of male faculty and much lower academic salaries.)

b. States with larger reported numbers of hate groups tend to have more people on death row. (Here, also tell how you could adjust for the lurking variable to uncover a more informative relationship.)

c. States with higher reported rates of gun ownership tend to have lower reported rates of violent crime.

**Interpreting $r^2$**

You might have noticed that computer outputs for regression analysis, like that in Display 3.18 on page 127, give the value of $R$-squared, or $r^2$, rather than the value of $r$. The student in this discussion will show you how to think about $r^2$ as the fraction of the variation in the values of $y$ that you can eliminate by taking $x$ into account.

Alexis: I've invented another way to measure the strength of the relationship between $x$ and $y$. It's based on the idea that the less variation there is from the linear trend, the stronger the correlation.

Statistician: How does it work?

Alexis: Let me ask the questions for a change. What is the best way to predict the values of $y$?

Statistician: I'd fit a least squares line and use the equation $\hat{y} = b_0 + b_1 x$.

Alexis: And how would you measure your total error?

Statistician: Well, I'd use the sum of the squares of the residuals, just like we did in the last section:

$$SSE = \sum (y - \hat{y})^2$$

Alexis: That's what I hoped you would say.

Statistician: We consultants try to be helpful.

Alexis: Don't get cocky. I'm about to change the rules. Pretend that you can't use the information about $x$. You don't have the $x$-values, and you want a single fitted value for $y$. What value would you choose?

Statistician: I'd use $\bar{y}$.

Alexis: And could you again use the sum of the squared errors to measure your total error?

Statistician: Sure, it's almost like the standard deviation. I'd find the sum of the squares of the deviations from the mean, which I'll call the total sum of squared error, SST:

$$SST = \sum (y - \bar{y})^2$$
Alexis: Okay, now for my new way to measure the strength of the relationship. If you have a strong relationship, the SSE will be small compared to the SST, right?

Statistician: Right.

Alexis: On the other hand, if you have a weak relationship, $x$ isn't much use for predicting $y$, and the SSE will be almost as big as the SST. Right?

Statistician: I see where you're going with this. You can use a ratio to measure the strength of the relationship.

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Strong</th>
<th>Weak</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSE</td>
<td>SST</td>
<td></td>
</tr>
</tbody>
</table>

Alexis: Exactly. Except now I have a problem. My ratio is near 0 when the relationship is strong, and it's near 1 when the relationship is weak. That's backward! Oh, I see how to fix it. Just subtract my ratio from 1.

$$1 - \frac{\text{SSE}}{\text{SST}} = \frac{\text{SST} - \text{SSE}}{\text{SST}}$$

Statistician: Good. Your new ratio is near 1 when the relationship is strong and near 0 when the relationship is weak. Your old ratio, SSE/SST, gave the proportion of error still there after the regression, so your new ratio . . .

Alexis: I can handle it from here. SST is the total error I started with. SST minus SSE is the amount of error I get rid of by using the relationship of $y$ with $x$. So my new ratio is the proportion of error I eliminate by using the regression.

Statistician: Right!

Alexis: But now I have two measures of strength—the correlation, $r$, and my new ratio. Which one should I use?

Statistician: Lucky for you—with a little algebra, they turn out to be equivalent.

$$r^2 = \frac{\text{SST} - \text{SSE}}{\text{SST}}$$

Alexis: Cool!

Statistician: We statisticians call $r^2$ the coefficient of determination. It tells us the proportion of variation in the $y$’s that is “explained” by $x$.

Alexis: I like it. Anything’s better than those $z$-scores!

**Predicting Pizzas**

Display 3.46 compares two sets of predictions of the calorie content in the seven kinds of pizza from E12 on page 134.
### Display 3.46  Two ways of predicting calorie content from fat content for seven kinds of pizza.

If you had to pick a single number as your predicted calorie content, you might choose the mean, 307.14 calories per serving. The fourth column in the table and the plot in Display 3.47 show the resulting errors.

<table>
<thead>
<tr>
<th>Fat (g)</th>
<th>Calories</th>
<th>Predicted</th>
<th>Error</th>
<th>Squared Error</th>
<th>Predicted</th>
<th>Error</th>
<th>Squared Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\hat{y}$</td>
<td>$(y - \hat{y})$</td>
<td>$(y - \hat{y})^2$</td>
<td>$\hat{y}$</td>
<td>$(y - \hat{y})$</td>
<td>$(y - \hat{y})^2$</td>
</tr>
<tr>
<td>9.0</td>
<td>230</td>
<td>307.14</td>
<td>−77.14</td>
<td>5,951.02</td>
<td>245.95</td>
<td>−15.95</td>
<td>254.42</td>
</tr>
<tr>
<td>19.5</td>
<td>385</td>
<td>307.14</td>
<td>77.86</td>
<td>6,061.73</td>
<td>402.66</td>
<td>−17.66</td>
<td>311.97</td>
</tr>
<tr>
<td>14.0</td>
<td>280</td>
<td>307.14</td>
<td>−27.14</td>
<td>736.74</td>
<td>320.58</td>
<td>−40.58</td>
<td>1,646.36</td>
</tr>
<tr>
<td>12.0</td>
<td>305</td>
<td>307.14</td>
<td>−2.14</td>
<td>4.59</td>
<td>290.73</td>
<td>14.27</td>
<td>203.76</td>
</tr>
<tr>
<td>8.0</td>
<td>230</td>
<td>307.14</td>
<td>−77.14</td>
<td>5,951.02</td>
<td>231.03</td>
<td>−1.03</td>
<td>1.05</td>
</tr>
<tr>
<td>14.2</td>
<td>350</td>
<td>307.14</td>
<td>42.86</td>
<td>1,836.73</td>
<td>323.56</td>
<td>26.44</td>
<td>699.06</td>
</tr>
<tr>
<td>15.0</td>
<td>370</td>
<td>307.14</td>
<td>62.86</td>
<td>3,951.02</td>
<td>335.50</td>
<td>34.50</td>
<td>1,190.23</td>
</tr>
<tr>
<td>Sum of Squared Errors</td>
<td>SST = 24,492.86</td>
<td>SSE = 4,306.85</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Display 3.47  Squared deviations around the mean of $y$.

If you use the regression equation, $\text{calories} = 112 + 14.9 \text{ fat}$, to predict calories, the resulting errors are much smaller in most cases and are given in the second to last column of Display 3.46 and shown on the plot in Display 3.48.

### Display 3.48  Squared deviations around the least squares line.
Using Alexis’s formula,

\[
\frac{\text{SST} - \text{SSE}}{\text{SST}} = \frac{24,492.86 - 4,306.85}{24,492.86} = 0.8242
\]

About 82% of the variation in calories among these brands of pizza can be attributed to fat content.

Taking the square root gives \( r = 0.908 \).

### DISCUSSION

#### Interpreting \( r^2 \)

D21. The scatterplot in Display 3.49 shows IQ plotted against head circumference, in centimeters, for a sample of 20 people. The mean IQ was 101, and the mean head circumference was 56.125 cm. The correlation is 0.138.

![Scatter Plot](image)

**Display 3.49**  

a. If you knew nothing about any possible relationship between head circumference and IQ, what IQ would you predict for a person with a head circumference of 54 cm?

b. The regression equation is \( IQ = 0.997 \cdot \text{head circumference} + 45 \). By about how much does this equation predict IQ will change with a 1-cm increase in head circumference?

c. What IQ does this equation predict for a person with a head circumference of 54 cm? How much faith do you have in this prediction?

d. How much of the variability in IQ is accounted for by the regression? Does the regression equation help you predict IQ in any practical sense?

e. If more people were added to the plot, how do you think the regression equation and correlation would change?

D22. Use Alexis’s formula for \( r^2 \) to explain why \(-1 \leq r \leq 1\).

D23. In a study of the effect of temperature on household heating bills, an investigator said, “Our research shows that about 70% of the variability in the number of heating units used by a particular house over the years can be explained by outside temperature.” Explain what the investigator might have meant by this statement.
Display 3.50 shows a hypothetical data set with the height of younger sisters plotted against the height of their older sisters. There is a moderate positive association: \( r = 0.337 \). For both younger and older sisters, the mean height is 65 in. and the standard deviation is 2.5 in. The line drawn on the first plot, \( y = x \), indicates the location of points representing the same height for both sisters. If you rotate your book and sight down the line, you can see that the points are scattered symmetrically about it.

In the second plot, look at the vertical strip for the older sisters with heights between 62 in. and 63 in. The \( X \) is at the mean height of the younger sisters with older sisters in this height range. It falls at about 64 in., not between 62 in. and 63 in. as you would expect. Looking at the vertical strip on the right, the mean height of younger sisters with older sisters between 68 in. and 69 in. is only about 66 in. If you were to use the line \( y = x \) to predict the height of the younger sister, you would tend to predict a height that is too small if the older sister is shorter than average and a height that is too large if the older sister is taller than average.

The flatter line through the third scatterplot in Display 3.50 is the least squares regression line. Notice that this line gets as close as it can to the center of each vertical strip. Thus, the least squares line is sometimes called the line of means. The predicted value of \( y \) at a given value of \( x \), using the regression line as the model, is the estimated mean of all responses that can be produced at that particular value of \( x \).

Notice that the regression line has a smaller slope than the major axis (\( y = x \)) of the ellipse. This means that the predicted values are closer to the mean than you might expect, which will always be the case for positively correlated data following a linear trend. The difference between these two lines is sometimes called the regression effect. If the correlation is near +1 or −1, the two lines will be nearly on top of each other and the regression effect will be minimal. For a moderate correlation such as that for the sisters’ heights, the regression effect will be quite large.

The regression effect was first noticed by British scientist Francis Galton around 1877. Galton noticed that the largest sweet-pea seeds tended to produce daughter seeds that were large but smaller than their parent. The smallest sweet-pea seeds tended to produce daughter seeds that were also small but larger than their parent. There was, in Galton’s words, a regression toward the mean. This is the origin of the term regression line. [Source: D. W. Forrest, *Francis Galton: The Life and Work of a Victorian Genius* (Taplinger, 1974).]

The regression effect is with us in everyday life whenever some element of chance is involved in a person’s score. For example, athletes are said to experience a “sophomore slump.” That is, athletes who have the best rookie seasons do not tend to be the same athletes who have the best second year. The top students on the second exam in your class probably did not do as well, relative to the rest of the class, on the first exam. The children of extremely tall or short parents do not tend to be as extreme in height as their parents. There does, indeed, seem to be a phenomenon at work that pulls us back toward the average. As Galton noticed, this prevents the spread in human height, for example, from increasing. Look for this effect as you work on regression analyses of data.
Display 3.50 Scatterplots showing the regression effect.
**Regression Toward the Mean**

D24. Why is the regression line sometimes called the “line of means”?

D25. The equation of the regression line for the scatterplot in Display 3.50

\[ y = 43.102 + 0.337x \]

Interpret the slope of this line in the context

of the situation and compare it to the interpretation of the slope of the

line \( y = x \).

**Summary 3.3: Correlation**

In your study of normal distributions in Chapter 2, you used the mean to tell

the center and then used the standard deviation as the overall measure of how

much the values deviated from that center. For "well-behaved" quantitative

relationships—that is, those whose scatterplots look elliptical—you use the

regression line as the center and then measure the overall amount of variation

from the line using the correlation, \( r \). You can think of the correlation, \( r \),

as the average product of the z-scores.

\[
    r = \frac{1}{n-1} \sum \left( \frac{x - \bar{x}}{s_x} \right) \left( \frac{y - \bar{y}}{s_y} \right) = \frac{1}{n-1} \sum z_x \cdot z_y
\]

Geometrically, the correlation measures how tightly packed the points of the

scatterplot are about the regression line.

- The correlation has no units and ranges from \(-1\) to \(+1\). It is unchanged if you

  interchange \( x \) and \( y \) or if you make a linear change of scale in \( x \) or \( y \), such as

  from feet to inches or from pounds to kilograms.

- In assessing correlation, begin by making a scatterplot and then follow these

  steps:

1. **Shape**: Is the plot linear, shaped roughly like an elliptical cloud, rather

   than curved, fan-shaped, or formed of separate clusters? If so, draw

   an ellipse to enclose the cloud of points. The data should be spread

   throughout the ellipse; otherwise, the pattern might not be linear or

   might have unusual features that require special handling. You should

   not calculate the correlation for patterns that are not linear.

2. **Trend**: If your ellipse tilts upward to the right, the correlation is

   positive; if it tilts downward to the right, the correlation is negative. The

   relationship between the correlation and the slope, \( b_1 \), of the regression

   line is given by

   \[
   b_1 = r \frac{s_y}{s_x}
   \]

3. **Strength**: If your ellipse is almost a circle or is horizontal, the relationship

   is weak and the correlation is near zero. If your ellipse is so thin that it

   looks like a line, the relationship is very strong and the correlation is

   near \(+1\) or \(-1\).
• Correlation is not the same as causation. Two variables may be highly correlated without one having any causal relationship with the other. The value of $r$ tells nothing about why $x$ and $y$ are related. In particular, a strong relationship between $x$ and $y$ might be due to a lurking variable.

You can interpret the value $r^2$ as the proportion of the total variation in $y$ that can be accounted for by using $x$ in the prediction model:

$$r^2 = \frac{\text{SST} - \text{SSE}}{\text{SST}}$$

The regression effect (or regression toward the mean) is the tendency of $y$-values to be closer to their mean than you might expect. That is, the regression line is flatter than the major axis of the ellipse surrounding the data.

### Practice

#### Estimating the Correlation

P9. By comparing to the plots in Display 3.41 on page 140, match each of the five scatterplots in Display 3.51 with its correlation, choosing from $-0.95$, $-0.5$, $0$, $0.5$, and $0.95$.

a.  

b.  

c.  

d.  

e.  

Display 3.51 Five scatterplots.

P10. The table in E12 (Display 3.31 on page 134) gives the amount of fat and number of calories in various pizzas.

a. Guess a value for the correlation, $r$.

b. Calculate $r$ using your calculator.

#### A Formula for the Correlation, $r$

P11. Eight artificial “data sets” are shown here. For each one, find the value of $r$, without computing if possible. Drawing a quick sketch might be helpful.

a.  

b.  

c.  

d.  

e.  

f.  

g.  

h.  

Display 3.51 Five scatterplots.
P12. The table in E12 (Display 3.31 on page 134) gives the amount of fat and number of calories in various pizzas. In P10, you used your calculator to find the correlation, \( r \). This time, make a table like that in Display 3.42 on page 142, and use the formula to find \( r \). What do you notice about the products \( z_x \cdot z_y \)?

P13. The scatterplot in Display 3.52 is divided into quadrants by vertical and horizontal lines that pass through the point of averages, \((\bar{x}, \bar{y})\).

\[ \begin{array}{c|c|c|c|c} \hline x & 0 & 1 & 2 & 3 \hline y & 0 & 1 & 2 & 3 \hline \end{array} \]

Display 3.52  Scatterplot divided into quadrants at the point of averages, \((\bar{x}, \bar{y})\).

a. Is the correlation positive or negative?

b. Give the coordinates of the point that will contribute the most to the correlation, \( r \).

c. Consider the product

\[
\frac{(x - \bar{x})(y - \bar{y})}{s_x s_y}
\]

Where are the points that have a positive product? How many of the 30 points have a positive product?

d. Where are the points that have a negative product? How many of the 30 points have a negative product?

Correlation and the Appropriateness of a Linear Model

P14. Both plots in Display 3.53 have a correlation of 0.26. For each plot, is fitting a regression line (as shown on the plot) an appropriate thing to do? Why or why not?

\[ \begin{array}{c|c|c|c|c} \hline x & 42 & 44 & 46 & 48 \hline y_1 & 0 & 5 & 10 & 15 \hline y_2 & 20 & 25 & 30 & 35 \hline \end{array} \]

Display 3.53  Two scatterplots with the same correlation.

The Relationship Between the Correlation and the Slope

P15. Imagine a scatterplot of two sets of exam scores for students in a statistics class. The score for a student on Exam 1 is graphed on the \( x \)-axis, and his or her score on Exam 2 is graphed on the \( y \)-axis. The slope of the regression line is 0.368. The mean of the Exam 1 scores is 72.99, and the standard deviation is 12.37. The mean of the Exam 2 scores is 75.80, and the standard deviation is 7.00.

a. Find the correlation of these scores.

b. Find the equation of the regression line for predicting an Exam 2 score from an Exam 1 score. Predict the Exam 2 score for a student who got a score of 80 on Exam 1.
c. Find the equation of the regression line for predicting an Exam 1 score from an Exam 2 score.

d. Sketch a scatterplot that could represent the situation described.

**Correlation Does Not Imply Causation**

P16. If you take a random sample of U.S. cities and measure the number of fast-food franchises in each city and the number of cases of stomach cancer per year in the city, you find a high correlation.

a. What is the lurking variable?
b. How would you adjust the data for the lurking variable to get a more meaningful comparison?

P17. If you take a random sample of public school students in grades K–12 and measure weekly allowance and size of vocabulary, you will find a strong relationship. Explain in terms of a lurking variable why you should not conclude that raising a student’s allowance will tend to increase his or her vocabulary.

P18. For the countries of the United Nations, there is a strong negative relationship between the number of TV sets per thousand people and the birthrate. What would be a careless conclusion about cause and effect? What is the lurking variable?

**Interpreting \( r^2 \)**

P19. Data on the association between high school graduation rates and the percentage of families living in poverty for the 50 U.S. states were presented in E26. Display 3.54 contains the scatterplot and a standard computer output of the regression analysis.

The regression equation is

\[
\text{Poverty} = 64.8 - 0.621 \text{HSG}
\]

**Display 3.54**  Poverty rates versus high school graduation rates.

a. Under “SOURCE,” the “Total” variation is the SST, and the “Error” variation is the SSE. From this information, find \( r \), the correlation.

b. Write an interpretation for \( r^2 \) in the context of these data.

c. Does the presence of a linear relationship here imply that a state that raises its graduation rate will cause its poverty rate to go down? Explain your reasoning.

d. What are the units for each of the values \( x, y, b_1 \), and \( r \)?
**Regression Toward the Mean**

P20. The plot in Display 3.55 shows the heights of older sisters plotted against the heights of their younger sisters. On a copy of this scatterplot, draw vertical lines to divide the points into six groups. Mark the approximate location of the mean of the $y$-values of each vertical strip. Sketch the regression line, $\hat{y} = 43 + 0.337x$. Note that the regression line comes as close as possible to the mean of each vertical strip. Now draw an ellipse around the data and connect the two ends of the ellipse. Is the regression line “flatter” than this line? Does this plot show the regression effect?

P21. Display 3.56 shows the first two exam scores for 29 college students enrolled in an introductory statistics course. Do you see any evidence of regression to the mean? If so, explain the nature of the evidence.

**Exercises**

E27. Each scatterplot in Display 3.57 was made on the same set of axes. Match each scatterplot with its correlation, choosing from $-0.06$, $0.25$, $0.40$, $0.52$, $0.66$, $0.74$, $0.85$, and $0.90$.

\[\begin{array}{cccc}
  \text{a.} & \text{b.} & \text{c.} & \text{d.} \\
  \end{array}\]

\[\begin{array}{cccc}
  \text{e.} & \text{f.} & \text{g.} & \text{h.} \\
  \end{array}\]

**Display 3.57** Eight scatterplots with various correlations.

E28. Estimate the correlation between the variables in these scatterplots.

\[\begin{array}{cccc}
  \text{a.} & \text{b.} & \text{c.} & \text{d.} \\
  \end{array}\]

b. The graduation rate versus the 75th percentile of SAT scores in E5 on page 113.

c. The college graduation rate versus the percentage of students in the top 10% of their high school graduating class in E5 on page 113.

E29. For each set of pairs, $(x, y)$, compute the correlation by hand, standardizing and finding the average product.

\[\begin{array}{cccc}
  \text{a.} & \text{b.} & \text{c.} & \text{d.} \\
  \end{array}\]

E30. For each artificial data set in P11 on page 155, compute the correlation by hand, standardizing and finding the average product.

E31. The scatterplot in Display 3.58 shows part of the hat size data of E6 on page 113. The plot is divided into quadrants by vertical and horizontal lines that pass through the point of averages, $(\bar{x}, \bar{y})$. 

**Display 3.55** The heights of older sisters versus the heights of their younger sisters.

**Display 3.56** Exam scores.
3.3 Correlation: The Strength of a Linear Trend

Display 3.58  Head circumference, in inches, versus hat size.

a. Estimate the value of the correlation.
b. Using the idea of standardized scores, explain why the correlation is positive.
c. Identify the point that contributes the most to the correlation. Explain why the contribution it makes is large.
d. Identify a point that contributes little to the correlation. Explain why the contribution it makes is small.

E32. The ellipses in Display 3.59 represent scatterplots that have a basic elliptical shape.

A. B. C.

Display 3.59  Three pairs of elliptical scatterplots.

a. Match these conditions with the corresponding pair of ellipses.
   I. One $s_y$ is larger than the other, the $s_x$’s are equal, and the correlations are strong.
   II. One of the correlations is stronger than the other, the $s_x$’s are equal, and the $s_y$’s are equal.
   III. One $s_x$ is larger than the other, the $s_y$’s are equal, and the correlations are weak.
b. Draw a pair of elliptical scatterplots to illustrate each comparison.
   i. One $s_y$ is larger than the other, the $s_x$’s are equal, and the correlations are weak.
   ii. One $s_x$ is larger than the other, the $s_y$’s are equal, and the correlations are strong.

E33. Several biology students are working together to calculate the correlation for the relationship between air temperature and how fast a cricket chirps. They all use the same crickets and temperatures, but some measure temperature in degrees Celsius and others measure it in degrees Fahrenheit. Some measure chirps per second, and others measure chirps per minute. Some use $x$ for temperature and $y$ for chirp rate, while others have it the other way around.

a. Will all the students get the same value for the slope of the least squares line? Explain why or why not.
b. Will they all get the same value for the correlation? Explain why or why not.

E34. For the sample of top-rated universities in E5 on page 113, the graduation rate has mean 82.7% and standard deviation 8.3%. The student/faculty ratio has mean 11.7 and standard deviation 4.3. The correlation is $-0.5$.

a. Find the equation of the least squares line for predicting graduation rate from student/faculty ratio.
b. Find the equation of the least squares line for predicting student/faculty ratio from graduation rate.
E35. These questions concern the relationship between the correlation, $r$, and the slope, $b_1$, of the regression line.

a. If $y$ is more variable than $x$, will the slope of the least squares line be greater (in absolute value) than the correlation? Justify your answer.

b. For a list of pairs $(x, y)$, $r = 0.8$, $b_1 = 1.6$, and the standard deviations of $x$ and $y$ are 25 and 50. (Not necessarily in that order.) Which is the standard deviation for $x$? Justify your answer.

c. Students in a statistics class estimated and then measured their head circumferences in inches. The actual circumferences had $SD = 0.93$, and the estimates had $SD = 4.12$. The equation of the least squares line for predicting estimated values from actual values was $\hat{y} = 11.97 + 0.36x$. What was the correlation?

d. What would be the slope of the least squares line for predicting actual head circumferences from the estimated values?

E36. Lost final exam. After teaching the same history course for about a hundred years, an instructor has found that the correlation, $r$, between the students' total number of points before the final examination and the number of points scored on their final examination is 0.8. The pre-final-exam point totals for all students in this year's course have mean 280 and $SD = 30$. The points on the final exam have mean 75 and $SD = 8$. The instructor's dog ate Julie's final exam, but the instructor knows that her total number of points before the exam was 300. He decides to predict her final exam score from her pre-final-exam total. What value will he get?

E37. Lurking variables. For each scenario, state a careless conclusion assuming cause and effect, and then identify a possible lurking variable.

a. For a large sample of different animal species, there is a strong positive correlation between average brain weight and average life span.

b. Over the last 30 years, there has been a strong positive correlation between the average price of a cheeseburger and the average tuition at private liberal arts colleges.

c. Over the last decade, there has been a strong positive correlation between the price of an average share of stock, as measured by the S&P 500, and the number of Web sites on the Internet.

E38. Manufacturers of low-fat foods often increase the salt content in order to keep the flavor acceptable to consumers. For a sample of different kinds and brands of cheeses, Consumer Reports measured several variables, including calorie content, fat content, saturated fat content, and sodium content. Using these four variables, you can form six pairs of variables, so there are six different correlations. These correlations turned out to be either about 0.95 or about $-0.5$.

a. List all six pairs of variables, and for each pair decide from the context whether the correlation is close to 0.95 or to $-0.5$.

b. State a careless conclusion based on taking the negative correlations as evidence of cause and effect.

c. Explain the negative correlation using the idea of a lurking variable.

E39. A study to determine whether ice cream consumption depends on the outside temperature gave the results shown in Display 3.60.
a. Use the values of SST and SSE in the regression analysis to compute \( r \), the correlation for the relationship between the temperature in degrees Fahrenheit and the number of pints of ice cream consumed per person. Check your answer against R-sq in the analysis.

b. Compute the value of the residual that is largest in absolute value.

c. Is there a cause-and-effect relationship between the two variables?

d. What are the units for each of \( x \), \( y \), \( b_1 \), and \( r \)?

e. The letters MS stand for “mean square.” How do you think the MS is computed?
E40. The scatterplot in Display 3.61 shows part of the aircraft data of Display 3.12 on page 115. For these data, \( r^2 = 0.83 \). Should \( r^2 \) be used as a statistical measure for these data? If so, interpret this value of \( r^2 \) in the context of the data. If not, explain why not.

Display 3.61 Scatterplot of number of seats versus fuel consumption (gal/h) for passenger aircraft.

E41. Suppose a teacher always praises students who score exceptionally well on a test and always scolds students who score exceptionally poorly. Use the notion of regression toward the mean to explain why the results will tend to suggest the false conclusion that scolding leads to improvement whereas praise leads to slacking off.

E42. A few years ago, a school in New Jersey tested all its 4th graders to select students for a program for the gifted. Two years later, the students were retested, and the school was shocked to find that the scores of the gifted students had dropped, whereas the scores of the other students had remained, on average, the same. What is a likely explanation for this disappointing development?

3.4 Diagnostics: Looking for Features That the Summaries Miss

As you learned in Chapter 2, summaries simplify. They are useful because they omit detail in order to emphasize a few general features. This quality also makes summaries potentially misleading, because sometimes the detail that is ignored has an important message to convey. Knowing just the mean and the standard deviation of a distribution doesn’t tell you if there are any outliers or whether the distribution is skewed. The same is true of the regression line and the correlation.

This section is about “diagnostics”—tools for looking beyond the summaries to see how well they describe the data and what features they leave out. The first part of this section deals with individual cases that stand apart from the overall pattern and with how these cases influence the regression line and the correlation. The second part shows you how to identify systematic patterns that involve many or all of the cases—the “shape” of the scatterplot.

Which Points Have the Influence?

Not all data points are created equal. You saw in the calculation of the correlation in Display 3.42 on page 142 that some points make large contributions and some small. Some make positive contributions and some negative. Your goal is to learn to recognize the points in a data set that might have an unusually large influence on where the regression line goes or on the size and sign of the correlation.
ACTIVITY 3.4 Near and Far

What you’ll need: an open area in which to step off distances

In this activity, you compare the actual distance to an object with what the distance appears to be.

1. Go to an open area, such as the hall or lawn of your school, and pick a spot as your origin. Choose six objects at various distances from the origin. Five of the objects should be within 10 to 20 paces, and the other should be a long way away (at least 100 paces).

2. For each of the six objects, estimate the number of paces from the origin to the object. Record your estimates.

3. From your origin, walk to each of your objects and count the actual number of regular paces it takes you to get there. Record this number beside your estimate.

4. Plot your data on a scatterplot, with your estimated value on the $x$-axis and the actual value on the $y$-axis. Does the plot show a linear trend?

5. Determine the equation of the regression line, and calculate the correlation.

6. Delete the point for the object that is farthest away from the origin. Determine the equation of the regression line and calculate the correlation for the reduced data set.

7. Did the extreme point have any influence on the regression line? On the correlation? Explain.

In Chapter 2, you learned about outliers for distributions—values that are separated from the bulk of the data. Outliers are atypical cases, and they can exert more than their share of influence on the mean and standard deviation. For scatterplots, as you will soon see, working with two variables together means that there can be outliers of various kinds. Different kinds of outliers can have different types of influence on the least squares line and the correlation. Unfortunately, there is no rule you can use to identify outliers in bivariate data. Just look for points surrounded by white space.

Judging a Point’s Influence

Points separated from the bulk of the data by white space are outliers and are potentially influential. To judge a point’s influence, compare the regression equation and correlation computed first with and then without the point in question.

To see these ideas in action, turn to the data on mammal longevity in Display 2.24 on page 43 and think about how to summarize the relationship between maximum and average longevity.
Example: Influential Mammals

The average elephant lives 35 years. The oldest elephant on record lived 70 years. The average hippo lives 41 years—longer than the average elephant—but the record-holding hippo lived only 54 years. The oldest-known beaver lived 50 years, almost as long as the champion hippo, but the average beaver cashes in his wood chips after only 5 short years of making them. Other mammals, however, are more predictable. If you look at the entire sample, shown in Display 3.62, it turns out that the elephant (E), hippo (H), and beaver (B) are the oddballs of the bunch. For the rest, there’s an almost linear relationship between average longevity and maximum longevity. The least squares line for the entire sample has the equation

\[ \hat{M} = 10.53 + 1.58A \]

where \( \hat{M} \), or “M-hat,” stands for predicted maximum longevity and \( A \) stands for observed average longevity. For every increase of 1 year in average longevity, the model predicts a 1.58-year increase in maximum longevity. The correlation for the relationship between these two variables is 0.77. How much influence do the oddballs have on these summaries?

![Maximum Longevity versus Average Longevity](image)

Display 3.62 Maximum longevity versus average longevity.

Solution

The hippo has the effect of pulling the right end of the regression line downward (like putting a heavy weight on one end of a seesaw), as you can see in Display 3.63. When the hippo is removed, that end of the regression line will “spring upward” and the slope will increase. Because one large residual has been removed and many of the remaining residuals have been reduced in size, the correlation will increase. The new slope is 1.96, and the new correlation is 0.80. The hippo has considerable influence on the slope and some influence on the correlation.

Now envision the scatterplot with just the elephant, E, missing. Because E is close to the straight line fit to the data, it produces a small residual. Thus, you would expect that removing E should not change the slope of the regression line much (not nearly as much as removing H did) and should reduce the correlation just a bit. In fact, the correlation does decrease some, to 0.72 from 0.77. However, the new slope is 1.53. It turns out that removing the elephant gives the hippo even more influence, and the slope decreases.
3.4 Diagnostics: Looking for Features That the Summaries Miss

Finally, envision the scatterplot with just the beaver, B, removed. B produces a large, positive residual close to the left end of the regression line. Thus, removing B should allow the left end of the line to drop, increasing the slope, and removing a large residual should increase the correlation. The new slope is 1.69 (an increase from 1.58), and the new correlation is 0.83 (an increase from 0.77). The beaver also has considerable influence on both slope and correlation.

Display 3.63 Regression lines for maximum longevity versus average longevity, with and without the hippo.
With a little practice, you often can anticipate the influence of certain points in a scatterplot, as in the previous example, but it is difficult to state general rules. The best rule is the one given in the box on page 163: Fit the line with and without the questionable point and see what happens. Then report all the results, with appropriate explanations.

**DISCUSSION**

**Why the Anscombe Data Sets Are Important**

Display 3.64 shows four scatterplots. These plots, known as “the Anscombe data” after their inventor, are arguably the most famous set of scatterplots in all of statistics. The questions that follow invite you to figure out why statistics books refer to them so often. In the process, you’ll learn more about what a summary doesn’t tell you about a data set.


D26. For each plot in Display 3.64, first give a short verbal description of the pattern in the plot. Then

a. either fit a line by eye and estimate its slope or tell why you think a line is not a good summary

b. either estimate the correlation by eye or tell why you think a correlation is not an appropriate summary

D27. Display 3.65 shows a computer output for one of the four Anscombe data set plots. Can you tell which one? If so, tell how you know. If not, explain why you can't tell.
Diagnosis: Looking for Features That the Summaries Miss

Dependent variable is: \( y \)

No Selector

\[ R^2 = 66.6\% \quad \text{R squared (adjusted)} = 62.9\% \]

\[ s = 1.297 \text{ with } 11 - 2 = 9 \text{ degrees of freedom} \]

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>27.5000</td>
<td>1</td>
<td>27.5000</td>
<td>18.0</td>
</tr>
<tr>
<td>Residual</td>
<td>13.7763</td>
<td>9</td>
<td>1.53070</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient s.e. of Coeff</th>
<th>t-ratio</th>
<th>prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>3.00091</td>
<td>1.125</td>
<td>2.67</td>
</tr>
<tr>
<td>( x )</td>
<td>0.500000</td>
<td>0.1180</td>
<td>4.24</td>
</tr>
</tbody>
</table>

Display 3.65  Regression analysis for one of the Anscombe data sets.


<table>
<thead>
<tr>
<th>Data Set I</th>
<th>Data Set II</th>
<th>Data Set III</th>
<th>Data Set IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>( y_1 )</td>
<td>( x_2 )</td>
<td>( y_2 )</td>
</tr>
<tr>
<td>10</td>
<td>8.04</td>
<td>10</td>
<td>9.14</td>
</tr>
<tr>
<td>8</td>
<td>6.95</td>
<td>8</td>
<td>8.14</td>
</tr>
<tr>
<td>13</td>
<td>7.58</td>
<td>13</td>
<td>8.74</td>
</tr>
<tr>
<td>9</td>
<td>8.81</td>
<td>9</td>
<td>8.77</td>
</tr>
<tr>
<td>11</td>
<td>8.33</td>
<td>11</td>
<td>9.26</td>
</tr>
<tr>
<td>14</td>
<td>9.96</td>
<td>14</td>
<td>8.10</td>
</tr>
<tr>
<td>6</td>
<td>7.24</td>
<td>6</td>
<td>6.13</td>
</tr>
<tr>
<td>4</td>
<td>4.26</td>
<td>4</td>
<td>3.10</td>
</tr>
<tr>
<td>12</td>
<td>10.84</td>
<td>12</td>
<td>9.13</td>
</tr>
<tr>
<td>7</td>
<td>4.82</td>
<td>7</td>
<td>7.26</td>
</tr>
<tr>
<td>5</td>
<td>5.68</td>
<td>5</td>
<td>4.74</td>
</tr>
</tbody>
</table>

Display 3.66  Anscombe plot data values.

a. Which plot has a point that is highly influential both with respect to the slope of the regression line and with respect to the correlation?

b. Compared to the other points in the plot, does the influential point lie far from the least squares line or close to it?

c. How would the slope and correlation change if you were to remove this point? Discuss this first without actually performing the calculations. Then carry out the calculations to verify your conjectures.

Residual Plots: Putting Your Data Under a Microscope

As you can see from the Anscombe plots, there are many features of the shape of a scatterplot that you can't learn from the standard set of summary numbers. Only when the cloud of points is elliptical, as in Display 3.41 on page 140, does the least squares line, together with the correlation, give a good summary of the relationship described by the plot. If the cloud of points isn't elliptical, these summaries aren't appropriate. How can you decide?
A special kind of scatterplot, called a residual plot, often can help you see more clearly what's going on. For some data sets, a residual plot can even show you patterns you might otherwise have overlooked completely. Statisticians use residual plots the way a doctor uses a microscope or an X ray—to get a better look at less obvious aspects of a situation. (Plots you use in this way are called “diagnostic plots” because of the parallel with medical diagnosis.) Push the analogy just a little. You’re the doctor, and data sets are your patients. Sets with elliptical clouds of points are the “healthy” ones; they don’t need special attention.

A **residual plot** is a scatterplot of residuals, \( y - \hat{y} \), versus predictor values, \( x \) (or, sometimes, versus predicted values, \( \hat{y} \)).

**Example: Constructing a Residual Plot**

Return to the data on percentage of on-time arrivals versus mishandled baggage for airlines, introduced in P2 on page 110. Calculate the residuals and make a plot.

**Solution**

Visualize each residual—the difference between the observed value of \( y \) and the predicted value, \( \hat{y} \)—as a vertical segment on the scatterplot in Display 3.67.

The calculated residuals are shown in Display 3.68, with the list of carriers ordered from smallest to largest on the \( x \)-scale. This allows the size of the residuals in the far right column to appear in the same order as in Display 3.67. Alaska produces a negative residual of modest size, whereas US Airways produces a large positive residual.

The residual plot, Display 3.69, is simply a scatterplot of the residuals versus the original \( x \)-variable, *mishandled baggage*. Note that 0 is at the middle of the residuals on the vertical scale.
### Display 3.68  Table showing residuals for the airline data.

<table>
<thead>
<tr>
<th>Airline</th>
<th>( x )</th>
<th>( y )</th>
<th>Predicted Value, ( \hat{y} )</th>
<th>Residual, ( y - \hat{y} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>United</td>
<td>4.00</td>
<td>80.9</td>
<td>78.5931</td>
<td>2.3069</td>
</tr>
<tr>
<td>America West</td>
<td>4.36</td>
<td>81.9</td>
<td>77.8182</td>
<td>4.0818</td>
</tr>
<tr>
<td>Southwest</td>
<td>4.42</td>
<td>78.4</td>
<td>77.6891</td>
<td>0.7109</td>
</tr>
<tr>
<td>Continental</td>
<td>4.62</td>
<td>75.7</td>
<td>77.2586</td>
<td>-1.5586</td>
</tr>
<tr>
<td>Northwest</td>
<td>5.36</td>
<td>67.2</td>
<td>75.6658</td>
<td>-8.4658</td>
</tr>
<tr>
<td>JetBlue</td>
<td>5.92</td>
<td>73.8</td>
<td>74.4604</td>
<td>-0.6604</td>
</tr>
<tr>
<td>American</td>
<td>6.50</td>
<td>73.1</td>
<td>73.2120</td>
<td>-0.1120</td>
</tr>
<tr>
<td>Alaska</td>
<td>7.02</td>
<td>69.1</td>
<td>72.0927</td>
<td>-2.9927</td>
</tr>
<tr>
<td>US Airways</td>
<td>7.16</td>
<td>78.3</td>
<td>71.7914</td>
<td>6.5086</td>
</tr>
<tr>
<td>Delta</td>
<td>8.03</td>
<td>70.1</td>
<td>69.9187</td>
<td>0.1813</td>
</tr>
</tbody>
</table>

The residual plot shows nearly random scatter, with no obvious trends. This is the ideal shape for a residual plot, because it indicates that a straight line is a reasonable model for the trend in the original data. [You can use your calculator to create residual plots. See Calculator Note 3I.]

### Display 3.69  Residual plot for the airline data.

D29. In Display 3.69, identify which residual belongs to Delta and which to Northwest.
D30. To see how residual plots magnify departures from the regression line, compare the Anscombe plots in Display 3.64 with Display 3.70, which shows the four corresponding residual plots in scrambled order.

Display 3.70 Residual plots for the four Anscombe data sets.

a. Match each of the original scatterplots in Display 3.64 with its corresponding residual plot in Display 3.70.

b. Describe the overall difference between the original scatterplots and the residual plots. What do the scatterplots show that the residual plots don't? What do the residual plots show that the scatterplots don't?

What to Look For in a Residual Plot

A careful data analyst always looks at a residual plot.

If the original cloud of points is elliptical, so that a line is an appropriate summary, the residual plot will look like a random scatter of points.

Use residual plots to check for systematic departures from constant slope (linear trend) and constant strength (same vertical spread). Look in particular for plots that are curved or fan-shaped. It's true that for data sets with only one predictor value (like those in this chapter), you often can get a good idea of what the residual plot will look like by carefully inspecting the original scatterplot. Once in a while, however, you get a surprise.

Example: Interpreting a Residual Plot

E19 on page 136 introduced data on median height versus age for young girls. Display 3.71 shows the scatterplot of these data, with the regression line. The overall average growth rate for the 12-year period is the slope of the regression line.
The plot looks nearly linear, but is a line a suitable model?

Display 3.71  Median height versus age for young girls.

Solution

The residual plot, shown in Display 3.72, quite dramatically reveals that the trend is not as linear as first imagined. The curvature in the residual plot mimics the curvature in the original scatterplot, which is harder to see. A line is not a good model for these data.

Display 3.72  Residual plot of median height versus age for young girls.

Statistical software often plots residuals against the predicted values, \( \hat{y} \), rather than against the predictor values, \( x \). For simple linear regression, both plots have exactly the same shape as long as the slope of the regression line is positive.

DISCUSSION  

Types of Residual Plots

D31. Display 3.73 shows a scatterplot and two residual plots for the “data set” consisting of these three ordered pairs \((x, y)\): \((0, 1), (1, 0), \) and \((2, 2)\). One residual plot plots residuals versus predictor values, \( x \), the sort of plot you
get from graphing calculators. The other plots residuals versus predicted (fitted) $y$-values, or $\hat{y}$, the sort of plot you get from computer software packages. Explain how the residual plots were produced and how you can tell which residual plot is which. The equation of the least squares line is $\hat{y} = 0.5 + 0.5x$.

Display 3.73 A scatterplot and two residual plots.

**Summary 3.4: Diagnostics: Looking for Features That the Summaries Miss**

For the simplest clouds of data points—elliptical in shape, with linear trend and no outliers—you can summarize all the main features of a scatterplot with just a few numbers, mainly the slope of the fitted line, $y$-intercept, and correlation. Not all plots are this simple, however, and a good statistician always does diagnostic checks for outliers and influential points and for departures from constant slope or constant strength.

- Points separated from the bulk of the data by white space are outliers and potentially influential.
- To judge a point’s influence, fit a line to the data and compute a regression equation and a correlation first with and then without the point in question. If the change in the regression equation and correlation is meaningful in your situation, report both sets of summary statistics.

For some data sets, a residual plot can show patterns you might otherwise overlook. A residual plot is a scatterplot of residuals, $y - \hat{y}$, versus predictor values, $x$. A residual plot also can be constructed as a scatterplot of residuals, $y - \hat{y}$, versus fitted values, $\hat{y}$. Use residual plots to check for systematic departures from linearity and for constant variability in $y$ across the values of $x$. If the data aren’t linear, the residual plot doesn’t look random. If the data have nonconstant variability, the residual plot is fan-shaped.

**Practice**

**Which Points Have the Influence?**

P22. The data in Display 3.74 show some interesting patterns in the relationship between domestic and international gross income from the ten movies with the highest domestic gross ticket sales.

a. Construct a scatterplot suitable for predicting international sales from domestic sales. Describe the pattern in the data.
b. Find the least squares line and the correlation for these data.
c. Remove the most influential data point and recalculate the least squares line and correlation. Describe the influence of the removed point.

<table>
<thead>
<tr>
<th>Movie</th>
<th>Domestic (U.S. $ millions)</th>
<th>International (U.S. $ millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Titanic</td>
<td>601</td>
<td>1235</td>
</tr>
<tr>
<td>Star Wars</td>
<td>461</td>
<td>337</td>
</tr>
<tr>
<td>Shrek 2</td>
<td>437</td>
<td>444</td>
</tr>
<tr>
<td>ET</td>
<td>435</td>
<td>322</td>
</tr>
<tr>
<td>Star Wars: The Phantom Menace</td>
<td>431</td>
<td>491</td>
</tr>
<tr>
<td>Pirates of the Caribbean: Dead Man’s Chest</td>
<td>417</td>
<td>492</td>
</tr>
<tr>
<td>Spider-Man</td>
<td>404</td>
<td>418</td>
</tr>
<tr>
<td>Star Wars: Revenge of the Sith</td>
<td>380</td>
<td>468</td>
</tr>
<tr>
<td>The Lord of the Rings: The Return of the King</td>
<td>377</td>
<td>752</td>
</tr>
<tr>
<td>Spider-Man 2</td>
<td>373</td>
<td>410</td>
</tr>
</tbody>
</table>

Display 3.74 Ticket sales for the ten highest-grossing domestic (United States and Canada) movies of all time. [Source: Internet Movie Database, us.imdb.com, September 12, 2006.]

P23. A data table and scatterplot of one student’s results from Activity 3.4a are shown in Display 3.75.

   a. How well did the student do in estimating the number of paces?
   b. Which point appears to be most influential?
   c. Calculate the slope of the regression line and the correlation with and without this point. Describe the influence of this point.

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>21</td>
<td>34</td>
</tr>
<tr>
<td>25</td>
<td>48</td>
</tr>
<tr>
<td>28</td>
<td>63</td>
</tr>
<tr>
<td>40</td>
<td>146</td>
</tr>
<tr>
<td>180</td>
<td>350</td>
</tr>
</tbody>
</table>

Display 3.75 Sample data from Activity 3.4a.

Residual Plots

P24. For the set of \((x, y)\) pairs \((0, 0), (0, 1), (1, 1),\) and \((3, 2)\), the equation of the least squares line is \(y = 0.5 + 0.5x.\)

   a. Plot the data and graph the least squares line.
   b. Next complete a table for the predicted values and residuals, like the table in Display 3.68 on page 169.
   c. Using the values in your table, plot residuals versus predictor, \(x.\)
   d. How does the residual plot differ from the scatterplot?
P25. Display 3.76 shows four scatterplots (A–D) for the data from a sample of commercial aircraft. Display 3.77 shows four corresponding residual plots (I–IV).

a. Match the residual plots to the scatterplots.

b. Using scatterplots A–D as examples, describe how you can identify each of these in a scatterplot from the residual plot.

i. a curve with increasing slope

ii. unequal variation in the responses

iii. a curve with decreasing slope

iv. two linear patterns with different slopes

c. For one of the plots, two line segments joined together seem to give a better fit than either a single line or a curve. Which plot is this? Is this pattern easier to see in the original scatterplot or in the residual plot?
E43. *Extreme temperatures.* The data in Display 3.78 provide the maximum and minimum temperatures ever recorded on each continent.

<table>
<thead>
<tr>
<th>Continent</th>
<th>Maximum Temperature (°F)</th>
<th>Minimum Temperature (°F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>136</td>
<td>−11</td>
</tr>
<tr>
<td>Antarctica</td>
<td>59</td>
<td>−129</td>
</tr>
<tr>
<td>Asia</td>
<td>129</td>
<td>−90</td>
</tr>
<tr>
<td>Australia</td>
<td>128</td>
<td>−9</td>
</tr>
<tr>
<td>Europe</td>
<td>122</td>
<td>−67</td>
</tr>
<tr>
<td>North America</td>
<td>134</td>
<td>−81</td>
</tr>
<tr>
<td>Oceania</td>
<td>108</td>
<td>12</td>
</tr>
<tr>
<td>South America</td>
<td>120</td>
<td>−27</td>
</tr>
</tbody>
</table>


a. Construct a scatterplot of the data suitable for predicting the minimum temperature from a given maximum temperature. Is a straight line a good model for these points? Explain.

b. Fit a least squares line to the points and calculate the correlation, even if you thought in part a that a straight line was not a good model.

c. Explain, in words and numbers, what influence Antarctica has on the slope of the regression line and on the correlation. How could an account of these data be misleading if it were not accompanied by a plot?

Two climbers stand on Mount Erebus, Antarctica, 12,500 ft above sea level.

E44. The data and plot in Display 3.79 are from E15 on page 135. They show the arsenic concentrations in the toenails of 21 people who used water from their private wells. Both measurements are in parts per million.

<table>
<thead>
<tr>
<th>Arsenic in Water (ppm)</th>
<th>Arsenic in Toenails (ppm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00087</td>
<td>0.119</td>
</tr>
<tr>
<td>0.00021</td>
<td>0.118</td>
</tr>
<tr>
<td>0</td>
<td>0.099</td>
</tr>
<tr>
<td>0.00115</td>
<td>0.118</td>
</tr>
<tr>
<td>0</td>
<td>0.277</td>
</tr>
<tr>
<td>0</td>
<td>0.358</td>
</tr>
<tr>
<td>0.00013</td>
<td>0.08</td>
</tr>
<tr>
<td>0.00069</td>
<td>0.158</td>
</tr>
<tr>
<td>0.00039</td>
<td>0.31</td>
</tr>
<tr>
<td>0</td>
<td>0.105</td>
</tr>
<tr>
<td>0</td>
<td>0.073</td>
</tr>
<tr>
<td>0.046</td>
<td>0.832</td>
</tr>
<tr>
<td>0.0194</td>
<td>0.517</td>
</tr>
<tr>
<td>0.137</td>
<td>2.252</td>
</tr>
<tr>
<td>0.0214</td>
<td>0.851</td>
</tr>
<tr>
<td>0.0175</td>
<td>0.269</td>
</tr>
<tr>
<td>0.0764</td>
<td>0.433</td>
</tr>
<tr>
<td>0</td>
<td>0.141</td>
</tr>
<tr>
<td>0.0165</td>
<td>0.275</td>
</tr>
<tr>
<td>0.00012</td>
<td>0.135</td>
</tr>
<tr>
<td>0.0041</td>
<td>0.175</td>
</tr>
</tbody>
</table>

Display 3.79 Arsenic concentrations.

a. Which point do you think has the most influence on the slope and correlation? What would be the effect of removing
this point? Perform the calculations to see if your intuition is correct.

b. Find a point that you think has almost no influence on the slope and correlation. Perform the calculations to see if your intuition is correct.

c. Find a point whose removal you think would make the correlation increase. Perform the calculations to see if your intuition is correct.

E45. How effective is a disinfectant? The data in Display 3.80 show (coded) bacteria colony counts on skin samples before and after a disinfectant is applied.

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
<th>Predicted</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>6</td>
<td>—?—</td>
<td>—?—</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>—?—</td>
<td>—?—</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>—?—</td>
<td>—?—</td>
</tr>
<tr>
<td>14</td>
<td>8</td>
<td>—?—</td>
<td>—?—</td>
</tr>
<tr>
<td>19</td>
<td>11</td>
<td>—?—</td>
<td>—?—</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>—?—</td>
<td>—?—</td>
</tr>
<tr>
<td>10</td>
<td>13</td>
<td>—?—</td>
<td>—?—</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>—?—</td>
<td>—?—</td>
</tr>
<tr>
<td>11</td>
<td>8</td>
<td>—?—</td>
<td>—?—</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>—?—</td>
<td>—?—</td>
</tr>
</tbody>
</table>

Display 3.80 Coded bacteria colony counts before (x) and after (y) treatment. [Source: Snedecor and Cochran, Statistical Methods (Iowa State University Press, 1967), p. 422.]

a. Plot the data, fit a regression line to them, and complete a copy of the table, filling in the predicted values and residuals.

b. Plot the residuals versus $x$, the count before the treatment. Comment on the pattern.

c. Use the residual plot to determine for which skin sample the disinfectant was unusually effective and for which skin sample it was not very effective.

E46. Textbook prices. Display 3.81 compares recent prices at a college bookstore to those of a large online bookstore.

<table>
<thead>
<tr>
<th>Type of Textbook</th>
<th>College Bookstore Price ($)</th>
<th>Online Bookstore Price ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemistry</td>
<td>93.40</td>
<td>94.18</td>
</tr>
<tr>
<td>Classic Fiction</td>
<td>9.95</td>
<td>7.96</td>
</tr>
<tr>
<td>English Anthology</td>
<td>46.70</td>
<td>48.75</td>
</tr>
<tr>
<td>Calculus</td>
<td>76.00</td>
<td>94.15</td>
</tr>
<tr>
<td>Biology</td>
<td>86.70</td>
<td>80.95</td>
</tr>
<tr>
<td>Statistics</td>
<td>7.95</td>
<td>6.36</td>
</tr>
<tr>
<td>Dictionary</td>
<td>24.00</td>
<td>16.80</td>
</tr>
<tr>
<td>Style Manual</td>
<td>12.70</td>
<td>10.66</td>
</tr>
<tr>
<td>Art History</td>
<td>66.00</td>
<td>45.50</td>
</tr>
</tbody>
</table>

Display 3.81 Prices for a sample of textbooks at a college bookstore and an online bookstore.

b. Construct a residual plot. Interpret it and point out any interesting features.

c. In comparing the prices of the textbooks, you might be more interested in a different line: $y = x$. Draw this line on a copy of the scatterplot in Display 3.81. What does it mean if a point lies above this line? Below it? On it?

d. A boxplot of the differences college price − online price is shown in Display 3.82. Interpret this boxplot.

Display 3.82 A boxplot of the differences between the college price and the online price for various textbooks.
E47. *Pizzas, again.* Display 3.83 shows the pizza data from E12 on page 134, with its regression line.

![Calories versus Fat, per 5-oz serving, for seven kinds of pizza.](image)

**Display 3.83** *Calories versus fat,* per 5-oz serving, for seven kinds of pizza.

a. Estimate the residuals from the graph, and use your estimates to sketch a rough version of a residual plot for this data set.

b. Which pizza has the largest positive residual? The largest negative residual? Are any of the residuals so extreme as to suggest that those pizzas should be regarded as exceptions?

c. Is any one of the pizzas a highly influential data point? If so, specify which one(s), and describe the effect on the slope of the fitted line and the correlation of removing the influential point or points from the analysis.

E48. *Aircraft.* Look again at Display 3.76 on page 174, which shows a scatterplot of flight length versus number of seats.

a. Does the slope of the pattern increase, decrease, or stay roughly constant as you move from left to right across the plot?

b. Focusing on the variation (spread) in flight length, \( y \), for planes with roughly the same seating capacity, compare the spreads for planes with few seats, a moderate number of seats, and a large number of seats. As you move from left to right across the plot, how does the spread change, if at all?

c. Suppose a friend chose a plane from the sample at random and told you the approximate number of seats. Could you guess its flight length to within 500 miles if the number of seats was between 50 and 150? If it was between 200 and 300? Explain.

d. What is the relationship between your answer in part b and residual plot I in Display 3.77?

e. Give an explanation for why the variation in flight length shows the pattern it does.

E49. Match each scatterplot (A–D) in Display 3.84 with its residual plot (I–IV) in Display 3.85. For which plots is a linear regression appropriate?
E50. Can either of the plots in Display 3.86 be a residual plot? Explain your reasoning.

A.

B.

Display 3.86 Residual Plots?

E51. Display 3.87 gives the data set for the three passenger jets from the example on page 123, along with a scatterplot showing the least squares line. (Values have been rounded.)

a. Use the equation of the line to find predicted values and residuals to complete the table in Display 3.87.

b. Use your numbers from part a to construct two residual plots, one with the predictor, $x$, on the horizontal axis and the other with the predicted value, $\hat{y}$, on the horizontal axis. How do the two plots differ?

<table>
<thead>
<tr>
<th>Aircraft</th>
<th>Seats</th>
<th>Cost</th>
<th>Predicted</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERJ-145</td>
<td>50</td>
<td>1100</td>
<td>—?—</td>
<td>—?—</td>
</tr>
<tr>
<td>DC-9</td>
<td>100</td>
<td>2100</td>
<td>—?—</td>
<td>—?—</td>
</tr>
<tr>
<td>MD-90</td>
<td>150</td>
<td>2700</td>
<td>—?—</td>
<td>—?—</td>
</tr>
</tbody>
</table>

Display 3.87 Cost per hour versus number of seats for three models of the passenger aircraft.

E52. Explain why a residual plot of $(x, \text{residual})$ and a plot of $(\text{predicted value}, \text{residual})$ have exactly the same shape if the slope of the regression line is positive. What changes if the slope is negative?

E53. Can you recapture the scatterplot from the residual plot? The residual plot in Display 3.88 was calculated from data showing the recommended weight (in pounds) for men at various heights over 64 in. The fitted weights ranged from 145 lb to 187 lb. Make a rough sketch of the scatterplot of these data.

E54. The plot in Display 3.89 shows the residuals resulting from fitting a line to the data for female life expectancy ($\text{life exp}$) versus gross national product (GNP, in thousands of dollars per capita) for a sample of countries from around the world. The regression equation for the sample data was

$$\text{life exp} = 67.00 + 0.63 \text{ GNP}$$

Sketch the scatterplot of $\text{life exp}$ versus GNP.

Display 3.89 Residuals of female life expectancy versus gross national product.
3.5 Shape-Changing Transformations

For scatterplots in which the points form an elliptical cloud, the regression line and correlation tell you pretty much all you need to know. But data don’t always behave so obligingly. For plots in which the points are curved, fan out, or contain outliers, the usual summaries do not tell you everything and can actually be misleading. What do you do then? This section shows you one possible remedy: Transform the data to get the shape you want.

You’re already familiar with linear transformations from Section 2.4—things like changing temperatures from degrees Fahrenheit to degrees Celsius or changing distances from feet to inches or times from minutes to seconds. These linear transformations—adding or subtracting a constant or multiplying or dividing by a constant—can change the center and spread of the distribution without changing its basic shape.

Nonlinear transformations, such as squaring each value or taking logarithms, do change the basic shape of the plot. This section shows how a transformation of a measurement scale can lead to simplified statistical analyses. One of the most common nonlinear relationships is the exponential, and that is where we begin our discussion.

Exponential Growth and Decay

Exponential functions often arise when you study how a quantity grows or decays as time passes. Many such quantities grow by an amount that is proportional to the amount present. This means that the amount present at one point in time is multiplied by a fixed constant to get the amount present at the next point in time, resulting in a function of the form \( y = ab^x \). The amount of growth of a population is proportional to population size, the growth of a bank account is proportional to the amount of money in the account, the amount by which a radioactive substance decays is proportional to how much is left, and the amount by which a cup of coffee cools is proportional to how much hotter it is than the air around it. For such situations, replacing \( y \) with \( \log y \) often gives a plot that is much more nearly linear than the original plot.

The examples in this section illustrate exponential growth and decay. At the same time, they illustrate another important feature of measurements taken over time: There is often an “up-and-down” pattern to the residuals. This pattern cannot be removed by a simple transformation.

Whenever your measurements come in chronological order, what happens next is likely to depend on what just happened. As a result, the patterns in your data might be subtler than the ones you’ve seen up to this point. In addition, the difference between a meaningful pattern and a quirk in the data might be harder to detect because the data typically show only one observation for a single point in time.

An Example of Exponential Decay

In Activity 3.5a, you will study a population that decreases exponentially over time.
**ACTIVITY 3.5a**

**Copper Flippers**

**What you’ll need:** 200 pennies, a paper cup

Count 200 pennies. You’ll use these for an exercise to see how fast pennies “die.”

A certain insect, the “copper flipper,” has a life span determined by the fact that there is a 50–50 chance a particular live flipper will die at the end of the day. So, on average, half of any population of copper flippers will die during the first day of life. Of those that survive the first day, on average half die during the second day, and so on.

By extraordinary coincidence, these bugs behave like a bunch of tossed pennies. If you toss 200 pennies, about half should come up heads and half tails. The heads represent the insects that survive the first day, and the tails represent those that die. You can collect the pennies that came up heads and toss them again to see how many survive the second day, and so on. This gives you a physical model for the distribution of the life span of the insects.

1. Place your pennies in a cup, shake them up, and toss them on a table. Count the number of heads, and record the number.
2. Set aside the pennies that came up tails. Place the pennies that came up heads back in the cup, and repeat the process.
3. After each toss (day), set the tails aside, collect the heads, count them, and toss them again. Repeat this process until you have fewer than five pennies left, but stop before you get to zero heads.
4. Construct a scatterplot of your data, with the number of the toss on the horizontal axis and the number of heads on the vertical axis. Does the pattern look linear?
5. How might you find an equation to summarize this pattern?

**Exponential Functions and Log Transformations**

How do you know if you have an exponential relationship of the form $y = ab^x$? One clue is that the points cluster about a function similar to those in Display 3.90. Another clue is that you have a variable whose values are mostly clustered at one end but range over two or more orders of magnitude (powers of 10). The best test, however, is to take the log of each value of $y$ and see if this will straighten the points.

Exponential relationships have an underlying model of the form $y = ab^x$. If $a > 1$, $a - 1$ gives the growth rate per time period. If $0 < a < 1$, $a - 1$ is negative and gives the decay rate. The points can be “linearized” (straightened) by taking the logarithm (base 10 or base $e$) of each value of $y$. The result will be a linear equation of the form

$$\log y = \log a + (\log b)x$$
Analyzing the Copper Flippers

Display 3.91 shows the scatterplot for one student’s data from Activity 3.5a.

Does a log transformation appear to be appropriate here? The pattern looks much like the left-hand curve in Display 3.90, and the values for the number of heads are clustered at the smaller values but range over two orders of magnitude. In addition, the number of heads remaining after each toss of the coins is roughly proportional to the number of coins tossed. A log transformation is worth a try. Display 3.92 shows the natural log (base $e$) of the number of heads plotted against the toss number, along with the regression line, for the data of Display 3.91.

Compare the scatterplot in Display 3.92 to the residual plot in Display 3.93. (The line segments are added to help your eye follow the time sequence.) Does the model appear to fit well if you look only at the scatterplot? How, if at all, does the residual plot alter your judgment of how well the line fits? The cyclical up-and-down pattern of residuals is common in such time series data.

The equation of the regression line (shown in Display 3.92) for the transformed data is given by the equation $\ln y = 5.21 - 0.66x$. If you solve this equation for $\hat{y}$, you get $\hat{y} = 183.1(0.52)^x$. The number of copper flippers that are alive each day is about 52% of the number alive the previous day. In other words, the decay rate is estimated to be 48%, or 0.48.
Chapter 3 Relationships Between Two Quantitative Variables

Toss Number

<table>
<thead>
<tr>
<th>Toss Number</th>
<th>ln(heads)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.5</td>
</tr>
<tr>
<td>2</td>
<td>1.5</td>
</tr>
<tr>
<td>3</td>
<td>0.5</td>
</tr>
<tr>
<td>4</td>
<td>1.5</td>
</tr>
<tr>
<td>5</td>
<td>2.5</td>
</tr>
<tr>
<td>6</td>
<td>3.5</td>
</tr>
</tbody>
</table>

Display 3.92  A plot of ln(heads) versus the number of the toss, with the regression line.

<table>
<thead>
<tr>
<th>Toss Number</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.375</td>
</tr>
<tr>
<td>2</td>
<td>-0.250</td>
</tr>
<tr>
<td>3</td>
<td>-0.125</td>
</tr>
<tr>
<td>4</td>
<td>0.000</td>
</tr>
<tr>
<td>5</td>
<td>0.125</td>
</tr>
<tr>
<td>6</td>
<td>0.250</td>
</tr>
</tbody>
</table>

Display 3.93  Residual plot for the ln(heads) regression.

Exponential Growth and Decay

D32. How would the scatterplot and the least squares line change if the coins had a probability of 0.6 of coming up heads? What insect death rate does this situation model?

An Example of Exponential Growth

Display 3.94 shows the population density (people per square mile) of the United States for all census years through 2000. For the years prior to 1960, only the 48 contiguous states are included. Alaska and Hawaii were added to the census in 1960. To find a reasonable model for this situation, start with a scatterplot.

Plot. Obviously, the pattern here is not linear. A curve of this type can be straightened by proportionally decreasing the large y-values (population densities, in this case). For variables like population growth (or the growth of many other phenomena), the logarithmic transformation works well. This transformation not only solves the data analysis problem nicely but also gives a neat interpretation to the resulting model.

Transform and Plot Again.  You can see the transformed points in Display 3.95.  For example, for the year 1800, the point (1800, ln 6.1) or (1800, 1.808) is plotted.

Fit.  Although there is still some curvature, the pattern is much more nearly linear, and a straight line might be a reasonable model to fit these data.  The regression line and regression analysis also are shown in Display 3.95 (on the next page).

The equation of the regression line is \( \hat{y} = -25.118 + 0.0148x \).  Solving for \( \hat{y} \) gives

\[
e^{\ln \hat{y}} = e^{-25.118 + 0.0148x}
\]

\[
\hat{y} = e^{-25.118} \left( e^{0.0148} \right)^x \text{ or } \hat{y} = e^{-25.118} (1.0149)^x
\]

This means that the population density is growing at about 1.5% per year.
The regression equation is

\[
\ln \text{density} = -25.118 + 0.0148 \text{ year}
\]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>Stdev</th>
<th>t-ratio</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-25.118</td>
<td>1.039</td>
<td>-24.18</td>
<td>0.000</td>
</tr>
<tr>
<td>year</td>
<td>0.0148405</td>
<td>0.0005479</td>
<td>27.09</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\[ s = 0.1630 \quad \text{R-sq} = 97.3\% \quad \text{R-sq(adj)} = 97.2\% \]

Analysis of Variance

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>19.502</td>
<td>19.502</td>
<td>733.64</td>
<td>0.000</td>
</tr>
<tr>
<td>Error</td>
<td>20</td>
<td>0.532</td>
<td>0.027</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>21</td>
<td>20.034</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Display 3.95  \( \ln (\text{population density}) \) versus year, with regression line and computer output.

**Residuals.** In the case of data over time, it is often advantageous to plot the residuals over time, as in Display 3.96.

Display 3.96  Residual plot of \( \ln (\text{population density}) \) versus year.

The result is not exactly random scatter! Well, it is about the best you can do. The problem is that there are subtle patterns in the data—and in the residuals—that no simple model will adequately account for.

**DISCUSSION**  

**Exponential Growth and Decay**

D33. Relate the pattern you see in the residual plot in Display 3.96 to the pattern of the data in the original scatterplot (Display 3.94).
D34. Why is there a huge jump from a large positive residual to a large negative residual as you move from 1800 to 1810? What events in U.S. history explain some of the other features of the residual plot?

D35. If you use a computer to fit a line to the \((year, density)\) data, it will automatically compute a correlation.
   a. Explain why a correlation is not a very useful summary for this data set.
   b. In Display 3.95, the computer gave the value of R-sq as 97.3% for the transformed data. Statisticians ordinarily are not very interested in the size of this diagnostic measure for time-ordered data. Can you think of any reasons why?

D36. What is the estimated annual rate of growth of the population density of the United States? What is the estimated rate of growth over a decade?

**Example: Logarithmic Transformation**

Do aircraft with a higher typical speed also have a greater average flight length? As you might expect, the answer is yes, but the relationship is nonlinear. (See Display 3.97.) Is there a simple equation that relates typical speed and flight length? The solution that follows leads you through one approach to these questions.

<table>
<thead>
<tr>
<th>Speed (mi/h)</th>
<th>Flight Length (mi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>520</td>
<td>3148</td>
</tr>
<tr>
<td>534</td>
<td>3960</td>
</tr>
<tr>
<td>494</td>
<td>2023</td>
</tr>
<tr>
<td>497</td>
<td>1637</td>
</tr>
<tr>
<td>495</td>
<td>1682</td>
</tr>
<tr>
<td>525</td>
<td>3515</td>
</tr>
<tr>
<td>509</td>
<td>3559</td>
</tr>
<tr>
<td>515</td>
<td>2485</td>
</tr>
<tr>
<td>560</td>
<td>947</td>
</tr>
<tr>
<td>472</td>
<td>1309</td>
</tr>
<tr>
<td>497</td>
<td>2122</td>
</tr>
<tr>
<td>464</td>
<td>1175</td>
</tr>
<tr>
<td>487</td>
<td>1987</td>
</tr>
<tr>
<td>454</td>
<td>1094</td>
</tr>
<tr>
<td>454</td>
<td>1035</td>
</tr>
<tr>
<td>446</td>
<td>886</td>
</tr>
<tr>
<td>430</td>
<td>644</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Speed (mi/h)</th>
<th>Flight Length (mi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>454</td>
<td>1065</td>
</tr>
<tr>
<td>409</td>
<td>646</td>
</tr>
<tr>
<td>432</td>
<td>791</td>
</tr>
<tr>
<td>441</td>
<td>879</td>
</tr>
<tr>
<td>403</td>
<td>542</td>
</tr>
<tr>
<td>442</td>
<td>904</td>
</tr>
<tr>
<td>396</td>
<td>465</td>
</tr>
<tr>
<td>339</td>
<td>175</td>
</tr>
<tr>
<td>407</td>
<td>576</td>
</tr>
<tr>
<td>387</td>
<td>496</td>
</tr>
<tr>
<td>398</td>
<td>587</td>
</tr>
<tr>
<td>387</td>
<td>313</td>
</tr>
<tr>
<td>360</td>
<td>343</td>
</tr>
<tr>
<td>397</td>
<td>486</td>
</tr>
<tr>
<td>357</td>
<td>382</td>
</tr>
<tr>
<td>230</td>
<td>202</td>
</tr>
</tbody>
</table>

**Display 3.97** Data table and scatterplot for the flight length and speed of various aircraft. ([Source: Air Transport Association of America, 2005, www.air-transport.org.])
Solution

The plot is curved much in the manner of exponential growth, so shrinking the $y$-scale is in order. Display 3.98 shows a scatterplot of $\ln(\text{flight length})$ versus $\text{speed}$, along with a least squares line and computer output.

![Graph of ln(flight length) versus speed](image)

The regression equation is

$$\ln(\text{Length}) = 1.57 + 0.0120 \times \text{speed}$$

The computer output is:

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>Stdev</th>
<th>t-ratio</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1.5730</td>
<td>0.3281</td>
<td>4.79</td>
<td>0.000</td>
</tr>
<tr>
<td>Speed</td>
<td>0.0119958</td>
<td>0.0007396</td>
<td>16.22</td>
<td>0.000</td>
</tr>
</tbody>
</table>

$s = 0.2691$  \hspace{0.5cm} R-sq = 89.5\%  \hspace{0.5cm} R-sq(adj) = 89.1\%$

Analysis of Variance

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>19.058</td>
<td>19.058</td>
<td>263.09</td>
<td>0.000</td>
</tr>
<tr>
<td>Error</td>
<td>31</td>
<td>2.246</td>
<td>0.072</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>32</td>
<td>21.303</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Display 3.98 \text{ln(flight length) versus speed.}

Although the pattern of points appears much more linear and the fit looks pretty good, the lack of randomness in the residual plot in Display 3.99 indicates that a linear model still does not really fit the points. D38 and E58 will give you a chance to continue the detective work.

![Residual plot](image)

Display 3.99 \text{Residual plot for ln(flight length) versus speed.}
Log Transformations

D37. How can you tell from Display 3.97 that the flight length values range over two orders of magnitude? Show how transforming from flight length to ln(flight length) will shrink the larger y-values more than the smaller ones and thus help straighten the plot.

D38. Describe the pattern in the residual plot in Display 3.99, and tell what it suggests as a next step in the analysis.

D39. In what sense, if any, is the relationship in Display 3.97 one of cause and effect? What is your evidence?

Log-Log Transformations of Power Functions

Wildlife biologists can estimate the length of an alligator without getting very close to it. However, to get its weight with any accuracy, they have to wrestle it onto a scale. This is a procedure that neither the biologist nor the alligator is happy to participate in.

Perhaps you can help the biologist predict the weight of an alligator spotted in the swamp from an estimate of its length. One way to do that is to collect data on both weight and length and then find a regression line that provides a good model of the relationship. Display 3.100 shows the weights and lengths of 25 alligators as measured by experts from the Florida Game and Freshwater Fish Commission.

<table>
<thead>
<tr>
<th>Length (in.)</th>
<th>Weight (lb)</th>
<th>Length (in.)</th>
<th>Weight (lb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>94</td>
<td>130</td>
<td>86</td>
<td>83</td>
</tr>
<tr>
<td>74</td>
<td>51</td>
<td>88</td>
<td>70</td>
</tr>
<tr>
<td>147</td>
<td>640</td>
<td>72</td>
<td>61</td>
</tr>
<tr>
<td>58</td>
<td>28</td>
<td>74</td>
<td>54</td>
</tr>
<tr>
<td>86</td>
<td>80</td>
<td>61</td>
<td>44</td>
</tr>
<tr>
<td>94</td>
<td>110</td>
<td>90</td>
<td>106</td>
</tr>
<tr>
<td>63</td>
<td>33</td>
<td>89</td>
<td>84</td>
</tr>
<tr>
<td>86</td>
<td>90</td>
<td>68</td>
<td>39</td>
</tr>
<tr>
<td>69</td>
<td>36</td>
<td>76</td>
<td>42</td>
</tr>
<tr>
<td>72</td>
<td>38</td>
<td>114</td>
<td>197</td>
</tr>
<tr>
<td>128</td>
<td>366</td>
<td>90</td>
<td>102</td>
</tr>
<tr>
<td>85</td>
<td>84</td>
<td>78</td>
<td>57</td>
</tr>
</tbody>
</table>

Display 3.100 Alligator weights and lengths. [Source: Florida Game and Freshwater Fish Commission.]

A scatterplot of the data (Display 3.101, on the next page) shows that the relationship is not linear. On thinking carefully about an appropriate model, you might realize that length is a linear measure while weight is related to volume—a cubic measure. So perhaps weight is related to the cube of length, or some power close to that. That is, the relationship is of the form weight = a \cdot length^3, where a is constant.
Power relationships have an equation of the form

\[ y = ax^b \]

as the underlying model. The points can be “linearized” (straightened) by taking the logarithm (base 10 or base e) of both the values of \( x \) and the values of \( y \). The result will be a linear equation of the form

\[ \log y = \log a + b \log x \]

Thus, if \( \ln(\text{weight}) \) is plotted against \( \ln(\text{length}) \), the plot should be fairly linear and the slope of the least squares line will provide an estimate of the power, \( b \). The result of this transformation is shown in Display 3.102. The regression equation is

\[ \ln(\text{weight}) = 3.29 \ln(\text{length}) - 10.2. \]

The plot does indeed look linear, and the estimate of \( b \) is 3.29. (Natural logs are used here, but logs to base 10 will produce essentially the same results.) The biologist can use this model to predict \( \ln(\text{weight}) \) from \( \ln(\text{length}) \) and then change the predicted value back to the original scale, if he or she chooses.

Note that the residual plot still shows a bit of curvature, mainly because the three largest alligators are somewhat influential. But the offsetting advantage of the power model is that the residuals are fairly homogeneous; that is, they don't tend to grow or shrink as length increases. This means that the error in the prediction of weight will be relatively constant for alligators of all lengths. The exponential model also fits these data well, but the residuals then lose their homogeneity with no substantial decrease in their size.

Ultimately, a model should be selected based on its intended use. These biologists wanted a model that predicts weight well for all reasonable values of length, not just for large alligators. Furthermore, a model should make sense to the experts in the field of use. The biologists could understand why weight (or volume) should have a cubic relationship with length but could see no reason why weight should grow exponentially with length.
Example: Using the Regression Equation

Use the regression equation to predict the weight of an alligator that is 100 inches in length.

Solution

The prediction is $\ln(\text{weight}) = 3.29\ln(100) - 10.2 \approx 4.951$. The predicted weight is found by solving

$$\ln(\text{weight}) \approx 4.951$$

$$\text{weight} = e^{4.951} \approx 141.3 \text{ lb}$$

Both biologist and alligator can rest more easily.

DISCUSSION

Log-Log Transformations of Power Functions

D40. For variable $x$ taking on the integer values from 1 through 10, sketch the graph of a power function with $a = 1$ and $b = 2$. Compare it to the exponential equation with $a = 1$ and $b = 2$. Discuss the differences between exponential models and power models.

D41. Use the alligator data to show that you get the same predicted weight for a given length whether you use base 10 or base $e$ logarithms.
Power Transformations

You always can use a log-log transformation to straighten points that follow a power relationship, \( y = ax^b \). However, sometimes your knowledge of a situation can allow you, with a little thought, to go directly to an appropriate power transformation without transforming through logarithms. For example, in exploring the relationship between time in free-fall and the distance an object falls, it is the square root of the distance that is linearly related to the time. In relating gas mileage to size of cars, miles per gallon could just as well be gallons per mile (a reciprocal transformation). Commonly used power transformations are given in the box.

To use a **power transformation**, transform \( y \) by replacing it with a power of \( y \), such as \( y^3 \), \( y^2 \), \( 1/y \), \( \sqrt{y} \), or \( \sqrt[3]{y} \).

The next example describes this process.

**Example: A Power Transformation**

The success of sustainable harvesting of timber depends on how fast trees grow, and one way to measure a tree’s growth rate is to find the relationship between its diameter and its age. If you know this relationship and the relationship between the age and height of a tree, then you can estimate the growth rate for total volume of timber. Displays 3.103 through 3.106 give data on the age (in years) and diameters (in inches at chest height) of a sample of oak trees, a scatterplot of diameter versus age, a residual plot, and numerical summaries in the form of computer output.

Based on these data, find a good model for predicting tree diameter from age.

<table>
<thead>
<tr>
<th>Age (yr)</th>
<th>Diameter (in.)</th>
<th>Age (yr)</th>
<th>Diameter (in.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.8</td>
<td>23</td>
<td>4.7</td>
</tr>
<tr>
<td>5</td>
<td>0.8</td>
<td>25</td>
<td>6.5</td>
</tr>
<tr>
<td>8</td>
<td>1.0</td>
<td>28</td>
<td>6.0</td>
</tr>
<tr>
<td>8</td>
<td>2.0</td>
<td>29</td>
<td>4.5</td>
</tr>
<tr>
<td>8</td>
<td>3.0</td>
<td>30</td>
<td>6.0</td>
</tr>
<tr>
<td>10</td>
<td>2.0</td>
<td>30</td>
<td>7.0</td>
</tr>
<tr>
<td>10</td>
<td>3.5</td>
<td>33</td>
<td>8.0</td>
</tr>
<tr>
<td>12</td>
<td>4.9</td>
<td>34</td>
<td>6.5</td>
</tr>
<tr>
<td>13</td>
<td>3.5</td>
<td>35</td>
<td>7.0</td>
</tr>
<tr>
<td>14</td>
<td>2.5</td>
<td>38</td>
<td>5.0</td>
</tr>
<tr>
<td>16</td>
<td>4.5</td>
<td>38</td>
<td>7.0</td>
</tr>
<tr>
<td>18</td>
<td>4.6</td>
<td>40</td>
<td>7.5</td>
</tr>
<tr>
<td>20</td>
<td>5.5</td>
<td>42</td>
<td>7.5</td>
</tr>
<tr>
<td>22</td>
<td>5.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.5 Shape-Changing Transformations

Display 3.104 Diameter versus age of oak trees, with regression line.

Display 3.105 Residual plot for the oak tree data.

The regression equation is
\[ \text{Diameter} = 1.15 + 0.163 \times \text{AGE} \]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>Stdev</th>
<th>t-ratio</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1.1507</td>
<td>0.4185</td>
<td>2.75</td>
<td>0.011</td>
</tr>
<tr>
<td>AGE</td>
<td>0.16278</td>
<td>0.01682</td>
<td>9.68</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\[ s = 1.021 \quad R\text{-sq} = 78.9\% \quad R\text{-sq(adj)} = 78.1\% \]

Analysis of Variance

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>97.593</td>
<td>97.593</td>
<td>93.63</td>
<td>0.000</td>
</tr>
<tr>
<td>Error</td>
<td>25</td>
<td>26.059</td>
<td>1.042</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>26</td>
<td>123.652</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Display 3.106 Numerical summary in the form of computer output of the ages and diameters (inches at chest height) of a sample of oak trees.

Solution

Inspection reveals that the point cloud is roughly elliptical with a slight downward curvature. Straightening this curve will require expanding the y-scale. You can formulate a power transformation by thinking carefully about the practical...
situation, rather than jumping immediately to the log-log transformation. If the diameter does not grow linearly with age, perhaps the cross-sectional area, \( A = \pi \left( \frac{d}{2} \right)^2 \), does. Try squaring the diameters.

Display 3.107 shows a scatterplot of \textit{diameter squared} versus \textit{age}, together with the equation of the least squares line. Display 3.108 shows the plot of \textit{residuals} versus \textit{age}. Even though the value of \( r^2 \) is about the same as before, the scatterplot and residual plot are now less curved. Taking all the evidence together, a linear model appears to fit the transformed data better than it does the original data.

![Display 3.107](image)

\textbf{Display 3.107} \textit{Diameter squared versus age} for the oak tree data, with the regression line.

![Display 3.108](image)

\textbf{Display 3.108} \textit{Residual plot} for \textit{diameter squared} versus \textit{age}.

Power transformations like the ones you have just seen can straighten a curved plot (or change a fan shape to a more nearly oval shape). By choosing the right power, you often can take a data set for which a fitted straight line and correlation are not suitable and convert it into one for which those summaries work well.

Is it cheating to change the shape of your data? (“You wanted a linear cloud, but you got a curved wedge. You didn’t like that, so you fiddled with the data until you got what you wanted.”) In fact, as you’ll see, changing scale is a matter of re-expressing the same data, not replacing the data with entirely new facts. The intelligent measurer selects a scale that is most useful for the problem the measurements were taken to solve.

[You can use your calculator to perform all the transformations you’ve learned about in this section. See \textit{Calculator Note 3J}.]
3.5 Shape-Changing Transformations

### Power Transformations

D42. Give a plausible explanation for why a tree might grow at a rate that makes the square of its diameter proportional to its age.

D43. If you need to predict the diameter of some oak trees for which you know only the age, would you rather do the predicting for 10-year-old trees or for 40-year-old trees? Explain your reasoning.

### Summary 3.5: Shape-Changing Transformations

For many scatterplots that show slopes or spreads that change as \( x \) changes, you can find a shape-changing transformation that brings your scatterplot much closer to the form for which the summaries of this chapter work best—a cloud of points in the shape of an ellipse with a linear trend. A transformation will not always make the pattern linear, however.

Transformations should have some basis in reality; they should not be simply chosen arbitrarily to see what might happen. Ideally, the transformation you use will be related in a plausible way to the situation that created your data.

The most common transformations are powers, in which you replace \( y \) with \( y^2 \), with \( \sqrt{y} \) or \( y^{1/2} \), with \( \frac{1}{y} \) or \( y^{-1} \), and so forth, and logarithms, in which you replace \( y \) with \( \log_{10} y \) or \( \ln y \). You can also transform \( x \).

Here are some other helpful facts to consider when choosing a transformation involving logarithms:

- A log-log transformation replacing \( x \) with \( \log x \) and \( y \) with \( \log y \) will straighten data modeled by a power function of the form \( y = ax^b \).
- Replacing \( y \) with \( \log y \) will straighten data modeled by an exponential function of the form \( y = ab^x \).
- If you have data on a quantity that changes over time by an amount roughly proportional to the quantity at a given time, then the logarithm of the quantity will be roughly a linear function of time.
- Consider a log transformation whenever you have a variable whose values are clustered at one end and range over two or more orders of magnitude (powers of 10).

For data collected over time (or over some other sequential ordering, such as distance along a path), there generally is one data value for each time, and each data value usually is correlated quite highly with the values to either side (its close neighbors). So the data will tend to have a much more intricate pattern than can be modeled by a straight line. Careful analysis of the pattern in the residuals often can help you see what is “really” happening over time.
Chapter 3 Relationships Between Two Quantitative Variables

Exponential Growth and Decay

P26. Dying dice. One of the authors gathered data on dying dice, starting with 200, and used the rule that a die “lives” if it lands showing 1, 2, 3, or 4. Here are the results:

200 122 81 58 29 19 11 8 6 4 2 2

a. Construct a scatterplot of the number of “live” dice versus the roll number.
b. Transform the number of live dice using natural logs, and construct a scatterplot of ln(dice) versus roll number.
c. Fit a line by the least squares method, and use its slope to estimate the rate of dying.
d. Plot residuals versus roll number, and describe the pattern.

P27. Florida is one of the fastest-growing states in the United States. The population figures for each census year from 1830 through 2000 are given in Display 3.109.

<table>
<thead>
<tr>
<th>Year</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>1830</td>
<td>34,730</td>
</tr>
<tr>
<td>1840</td>
<td>54,477</td>
</tr>
<tr>
<td>1850</td>
<td>87,445</td>
</tr>
<tr>
<td>1860</td>
<td>140,424</td>
</tr>
<tr>
<td>1870</td>
<td>187,748</td>
</tr>
<tr>
<td>1880</td>
<td>269,493</td>
</tr>
<tr>
<td>1890</td>
<td>391,422</td>
</tr>
<tr>
<td>1900</td>
<td>528,542</td>
</tr>
<tr>
<td>1910</td>
<td>752,619</td>
</tr>
<tr>
<td>1920</td>
<td>968,470</td>
</tr>
<tr>
<td>1930</td>
<td>1,468,211</td>
</tr>
<tr>
<td>1940</td>
<td>1,897,414</td>
</tr>
<tr>
<td>1950</td>
<td>2,771,605</td>
</tr>
<tr>
<td>1960</td>
<td>4,951,560</td>
</tr>
<tr>
<td>1970</td>
<td>6,791,418</td>
</tr>
<tr>
<td>1980</td>
<td>9,746,324</td>
</tr>
<tr>
<td>1990</td>
<td>12,937,926</td>
</tr>
<tr>
<td>2000</td>
<td>15,982,378</td>
</tr>
</tbody>
</table>


Exponential Functions and Log Transformations

P28. The logarithm of a number is the exponent when you write the number in the form “base raised to a power.” Thus, \( \log_{10} 1000 = 3 \) means the same thing as \( 10^3 \). Use this fact to transform \( y \) into logs (in base 10) in this set of \( (x, y) \) pairs: (2, 1000), (1, 100), (0, 10), and \((-1, 1)\). Then plot \( \log_{10} y \) versus \( x \) and check that the points lie in a straight line. Find the slope and \( y \)-intercept of the line.

P29. Repeat P28 for these two sets of pairs.
a. (6, 1000), (4, 100), (2, 10), (0, 1)
b. (5, 0.0001), (6, 0.01), (8, 100)
P30. Verify that if \( \log_{10} y = c + dx \), then \( y = ab^x \), where \( a = 10^c \) and \( b = 10^d \). Use this fact to rewrite each of your fitted equations in P28 and P29 in the form \( y = ab^x \).

P31. Rewrite the fitted equation in Display 3.98 on page 186 in the form

\[
\frac{\text{flight length}}{\text{speed}} = ab^{\text{speed}}
\]

P32. In setting standards for the consumption of fish tainted by chemicals in the water from which they were taken, the U.S. government commissioned a study of fish consumption for one such contaminated area. A part of the study involved interviews with a sample of noncommercial fishers that asked, among other things, how often the person fished in this water over the past month and how many fish meals his or her family consumed over the past month. The number of meals was then converted to grams of fish consumed (a statistical process in itself). The number of fishing trips is fairly easy for people to remember, but they aren’t so accurate at reporting the amount of fish their family ate. A good model relating the number of fishing trips to consumption would be helpful in estimating fish consumption. From the set of data in Display 3.110, find such a model.

<table>
<thead>
<tr>
<th>Trips</th>
<th>Consumption (g)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>17.2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>9.1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>13.15</td>
</tr>
<tr>
<td>3</td>
<td>9.1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>31.375</td>
</tr>
<tr>
<td>7</td>
<td>9.1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>25.3</td>
</tr>
<tr>
<td>7</td>
<td>9.1</td>
</tr>
<tr>
<td>1</td>
<td>5.05</td>
</tr>
<tr>
<td>5</td>
<td>13.15</td>
</tr>
<tr>
<td>29</td>
<td>82</td>
</tr>
<tr>
<td>5</td>
<td>11.125</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>3.025</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>13.15</td>
</tr>
</tbody>
</table>

Display 3.110  Fish consumption data.

Log-Log Transformations of Power Functions

P33. Use the regression equation on page 189 to predict the weight of an alligator that is 75 inches long.
P34. Having a good measure of tidal velocity (the speed at which water depth increases) in an estuary is critically important, especially during storms. Tidal velocity is difficult to measure, but it is related to the depth of the water. Thus, a good model of this relationship would allow scientists to predict the velocity from measurements of water depth. Display 3.111 shows measurements of the depth of water (in meters) and tidal velocities (in meters per second) for certain locations in an estuary.

<table>
<thead>
<tr>
<th>Depth (m)</th>
<th>Velocity (m/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02</td>
<td>0.6</td>
</tr>
<tr>
<td>0.02</td>
<td>0.55</td>
</tr>
<tr>
<td>0.2</td>
<td>0.88</td>
</tr>
<tr>
<td>0.2</td>
<td>0.85</td>
</tr>
<tr>
<td>0.4</td>
<td>1.1</td>
</tr>
<tr>
<td>0.4</td>
<td>1.05</td>
</tr>
<tr>
<td>0.6</td>
<td>1.0</td>
</tr>
<tr>
<td>0.6</td>
<td>1.2</td>
</tr>
<tr>
<td>0.8</td>
<td>1.15</td>
</tr>
<tr>
<td>0.8</td>
<td>0.95</td>
</tr>
<tr>
<td>1.0</td>
<td>1.2</td>
</tr>
<tr>
<td>1.0</td>
<td>1.05</td>
</tr>
</tbody>
</table>

Display 3.111  Tidal velocity versus the water depth.  
[Source: Shi et al., International Journal of Numerical Methods for Fluids, 2003.]

a. Describe the nature of the relationship between depth and velocity.

b. Fit an appropriate model that would allow the prediction of velocity from depth.

Power Transformations

P35. Three sets of pairs \((x, y)\) are given here. For each set, (i) plot \(y\) versus \(x\), (ii) find a power of \(y\) to use in place of \(y\) itself to get a linear plot, and (iii) plot that power of \(y\) versus \(x\).

a. \((0, 0), (1, 1), (2, 8), (3, 27)\)

b. \((1, 10), (2, 5), (5, 2), (10, 1)\)

c. \((100, 10), (64, 8), (25, 5), (9, 3), (1, 1)\)

P36. In P35, suppose that, instead of plotting a power of \(y\) versus \(x\), you plot \(y\) versus a power of \(x\). For each of the three sets of pairs \((x, y)\), find the power of \(x\) for which the points lie on a line. What is the relationship between the powers of \(y\) in P35 and the powers of \(x\) in this problem?

P37. For each of these relationships, first write the equation that relates \(x\) and \(y\). Then use this equation to find a power of \(y\) that you could plot against \(x\) in order to get a linear plot.

a. \(y\) is the area of a circle, and \(x\) is the radius of the circle.

b. \(y\) is the volume of a block whose sides all have equal lengths, and \(x\) is the side length.

c. \(y\) is the volume of an 8-ft section of log with a circular cross section, and \(x\) is the diameter of the log's cross section.

P38. If the square of a tree’s diameter is roughly proportional to its age, then you can expect the diameter itself to be roughly proportional to the square root of the tree’s age. For the data in Display 3.103 on page 190, use a computer or calculator to make a scatterplot of diameter versus square root of age, fit a least squares line, and plot the residuals. Which residual plot shows more of a fan shape, the plot of diameter squared versus age or the plot of diameter versus square root of age? If you want a plot that shows an elliptical cloud, which transformation should you choose?

P39. Display 3.112 shows the brain weights and body weights for a collection of mammals. The goal is to establish the relationship of brain weight to body weight.
a. Assuming that there is a power relationship here, can you guess what it is from the scatterplot? If \( y \) is written as a function of \( x \) to some power, should the power be greater than 1 or less than 1?

b. Plot \( \log(\text{brain}) \) versus \( \log(\text{body}) \). Describe the pattern of the plot.

c. Fit a line to the plot in part b. Write an equation relating \( y \) to \( x \). Does your equation support your answer to part a?

<table>
<thead>
<tr>
<th>Species</th>
<th>Brain Weight (g)</th>
<th>Body Weight (kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>African elephant</td>
<td>5712</td>
<td>6654</td>
</tr>
<tr>
<td>African giant pouched rat</td>
<td>6.6</td>
<td>1</td>
</tr>
<tr>
<td>Arctic fox</td>
<td>44.5</td>
<td>3.385</td>
</tr>
<tr>
<td>Arctic ground squirrel</td>
<td>5.7</td>
<td>0.92</td>
</tr>
<tr>
<td>Asian elephant</td>
<td>4603</td>
<td>2547</td>
</tr>
<tr>
<td>Baboon</td>
<td>179.5</td>
<td>10.55</td>
</tr>
<tr>
<td>Big brown bat</td>
<td>0.3</td>
<td>0.023</td>
</tr>
<tr>
<td>Cat</td>
<td>25.6</td>
<td>3.3</td>
</tr>
<tr>
<td>Chimpanzee</td>
<td>440</td>
<td>52.16</td>
</tr>
<tr>
<td>Chinchilla</td>
<td>6.4</td>
<td>0.425</td>
</tr>
<tr>
<td>Cow</td>
<td>423</td>
<td>465</td>
</tr>
<tr>
<td>Desert hedgehog</td>
<td>2.4</td>
<td>0.55</td>
</tr>
<tr>
<td>Donkey</td>
<td>419</td>
<td>187.1</td>
</tr>
<tr>
<td>Eastern American mole</td>
<td>1.2</td>
<td>0.075</td>
</tr>
<tr>
<td>European hedgehog</td>
<td>3.5</td>
<td>0.785</td>
</tr>
<tr>
<td>Giant armadillo</td>
<td>81</td>
<td>60</td>
</tr>
<tr>
<td>Giraffe</td>
<td>680</td>
<td>529</td>
</tr>
<tr>
<td>Goat</td>
<td>115</td>
<td>27.66</td>
</tr>
<tr>
<td>Golden hamster</td>
<td>1</td>
<td>0.12</td>
</tr>
<tr>
<td>Gorilla</td>
<td>406</td>
<td>207</td>
</tr>
<tr>
<td>Gray seal</td>
<td>325</td>
<td>85</td>
</tr>
<tr>
<td>Gray wolf</td>
<td>119.5</td>
<td>36.33</td>
</tr>
<tr>
<td>Ground squirrel</td>
<td>4</td>
<td>0.101</td>
</tr>
<tr>
<td>Guinea pig</td>
<td>5.5</td>
<td>1.04</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Species</th>
<th>Brain Weight (g)</th>
<th>Body Weight (kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Horse</td>
<td>655</td>
<td>521</td>
</tr>
<tr>
<td>Human</td>
<td>1320</td>
<td>62</td>
</tr>
<tr>
<td>Jaguar</td>
<td>157</td>
<td>100</td>
</tr>
<tr>
<td>Kangaroo</td>
<td>56</td>
<td>35</td>
</tr>
<tr>
<td>Lesser short-tailed shrew</td>
<td>0.14</td>
<td>0.005</td>
</tr>
<tr>
<td>Little brown bat</td>
<td>0.25</td>
<td>0.01</td>
</tr>
<tr>
<td>Mole rat</td>
<td>3</td>
<td>0.122</td>
</tr>
<tr>
<td>Mountain beaver</td>
<td>8.1</td>
<td>1.35</td>
</tr>
<tr>
<td>Mouse</td>
<td>0.4</td>
<td>0.023</td>
</tr>
<tr>
<td>Musk shrew</td>
<td>0.33</td>
<td>0.048</td>
</tr>
<tr>
<td>Nine-banded armadillo</td>
<td>10.8</td>
<td>3.5</td>
</tr>
<tr>
<td>North American opossum</td>
<td>6.3</td>
<td>1.7</td>
</tr>
<tr>
<td>Owl monkey</td>
<td>15.5</td>
<td>0.48</td>
</tr>
<tr>
<td>Pig</td>
<td>180</td>
<td>192</td>
</tr>
<tr>
<td>Rabbit</td>
<td>12.1</td>
<td>2.5</td>
</tr>
<tr>
<td>Raccoon</td>
<td>39.2</td>
<td>4.288</td>
</tr>
<tr>
<td>Rat</td>
<td>1.9</td>
<td>0.28</td>
</tr>
<tr>
<td>Red fox</td>
<td>50.4</td>
<td>4.235</td>
</tr>
<tr>
<td>Rhesus monkey</td>
<td>179</td>
<td>6.8</td>
</tr>
<tr>
<td>Roe deer</td>
<td>98.2</td>
<td>14.83</td>
</tr>
<tr>
<td>Sheep</td>
<td>175</td>
<td>55.5</td>
</tr>
<tr>
<td>Tree shrew</td>
<td>2.5</td>
<td>0.104</td>
</tr>
<tr>
<td>Water opossum</td>
<td>3.9</td>
<td>3.5</td>
</tr>
<tr>
<td>Yellow-bellied marmot</td>
<td>17</td>
<td>4.05</td>
</tr>
</tbody>
</table>

Exercises

E55. For the data in Display 3.97 on page 185, try fitting a straight line to the square root of flight length as a function of speed. Does this transformation work as well as the log transformation? Explain your reasoning.

E56. More dying dice. Follow the same steps as in P26 on page 194 for these numbers of surviving dice: 200, 72, 28, 9, 5, 2, and 1. Use your data to estimate what the probability of "dying" was in order to generate these numbers.

E57. Growing kids. Median heights and weights of growing boys are presented in Display 3.113. What model would you choose to predict weight from a boy’s known height?

<table>
<thead>
<tr>
<th>Age (yr)</th>
<th>Height (in.)</th>
<th>Weight (lb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>35.8</td>
<td>28.03</td>
</tr>
<tr>
<td>3</td>
<td>39.1</td>
<td>31.68</td>
</tr>
<tr>
<td>4</td>
<td>41.4</td>
<td>35.90</td>
</tr>
<tr>
<td>5</td>
<td>44.2</td>
<td>40.66</td>
</tr>
<tr>
<td>6</td>
<td>46.8</td>
<td>45.72</td>
</tr>
<tr>
<td>7</td>
<td>49.6</td>
<td>50.97</td>
</tr>
<tr>
<td>8</td>
<td>51.7</td>
<td>56.65</td>
</tr>
<tr>
<td>9</td>
<td>54.1</td>
<td>63.10</td>
</tr>
<tr>
<td>10</td>
<td>56.3</td>
<td>70.60</td>
</tr>
<tr>
<td>11</td>
<td>58</td>
<td>79.35</td>
</tr>
<tr>
<td>12</td>
<td>60.8</td>
<td>89.47</td>
</tr>
<tr>
<td>13</td>
<td>63.7</td>
<td>100.78</td>
</tr>
<tr>
<td>14</td>
<td>66.6</td>
<td>112.71</td>
</tr>
</tbody>
</table>


E58. Cost per seat per mile and flight length, revisited. As you saw in P25 on page 174, when cost per seat per mile is plotted against flight length, the pattern is not linear. The residual plot in Display 3.77 on page 174 strongly suggests that two line segments might provide a better model than a single line. Apparently, there is one relationship for aircraft meant for longer routes and another for aircraft meant for shorter routes. Look through the complete listing of the data (Display 3.12 on page 115) and the scatterplots from P25 to see whether any features other than flight length separate the aircraft into the same two groups.

E59. Chimp hunting parties. After Jane Goodall discovered that chimpanzees are not solely vegetarian, much research began into the behavior of chimpanzees as hunters. Some animals hunt alone or in small groups, while others hunt in large groups. Where does the chimp fit in, and what is the success rate of chimps’ hunting parties? Not surprisingly, the success of the hunt depends in part on the size of the hunting party. Display 3.114 gives some data on the number of chimps in a hunting party and the success rate of parties of that size.

<table>
<thead>
<tr>
<th>Number of Chimps</th>
<th>Percentage Successful</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td>3</td>
<td>28</td>
</tr>
<tr>
<td>4</td>
<td>42</td>
</tr>
<tr>
<td>5</td>
<td>40</td>
</tr>
<tr>
<td>6</td>
<td>58</td>
</tr>
<tr>
<td>7</td>
<td>45</td>
</tr>
<tr>
<td>8</td>
<td>62</td>
</tr>
<tr>
<td>9</td>
<td>65</td>
</tr>
<tr>
<td>10</td>
<td>63</td>
</tr>
<tr>
<td>12</td>
<td>75</td>
</tr>
<tr>
<td>13</td>
<td>75</td>
</tr>
<tr>
<td>14</td>
<td>78</td>
</tr>
<tr>
<td>15</td>
<td>75</td>
</tr>
<tr>
<td>16</td>
<td>82</td>
</tr>
</tbody>
</table>


a. Plot the data in a way that allows the building of a model to predict success from size of hunting party. Describe the pattern you see.
b. Will a simple linear regression model work well here? Why or why not?

c. Look for a transformation that will produce a model with better predicting ability than the simple linear one. Fit the model to the data.

d. Investigate the residuals from the model in part c. Are you happy with the fit of that model?

E60. The data in Display 3.115 are the population of the United States from 1830 through 2000 and the number of immigrants entering the country in the decade preceding the given year.

<table>
<thead>
<tr>
<th>Year</th>
<th>U.S. Population</th>
<th>Immigration (in thousands)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1830</td>
<td>12,866,020</td>
<td>152</td>
</tr>
<tr>
<td>1840</td>
<td>17,069,453</td>
<td>599</td>
</tr>
<tr>
<td>1850</td>
<td>23,191,876</td>
<td>1713</td>
</tr>
<tr>
<td>1860</td>
<td>31,433,321</td>
<td>2598</td>
</tr>
<tr>
<td>1870</td>
<td>39,818,449</td>
<td>2315</td>
</tr>
<tr>
<td>1880</td>
<td>50,155,783</td>
<td>1812</td>
</tr>
<tr>
<td>1890</td>
<td>62,947,714</td>
<td>5247</td>
</tr>
<tr>
<td>1900</td>
<td>75,994,575</td>
<td>3688</td>
</tr>
<tr>
<td>1910</td>
<td>91,972,266</td>
<td>8795</td>
</tr>
<tr>
<td>1920</td>
<td>105,710,620</td>
<td>5736</td>
</tr>
<tr>
<td>1930</td>
<td>122,775,046</td>
<td>4107</td>
</tr>
<tr>
<td>1940</td>
<td>131,669,275</td>
<td>528</td>
</tr>
<tr>
<td>1950</td>
<td>150,697,361</td>
<td>1035</td>
</tr>
<tr>
<td>1960</td>
<td>179,323,175</td>
<td>2515</td>
</tr>
<tr>
<td>1970</td>
<td>203,302,031</td>
<td>3322</td>
</tr>
<tr>
<td>1980</td>
<td>226,542,199</td>
<td>4493</td>
</tr>
<tr>
<td>1990</td>
<td>248,709,873</td>
<td>7338</td>
</tr>
<tr>
<td>2000</td>
<td>281,421,906</td>
<td>9095</td>
</tr>
</tbody>
</table>


a. Find the population growth for each decade. Was the increase in population constant from decade to decade? How would you describe the pattern?

b. Fit a model to the (year, population) data and defend your model as representative of the major trend(s) in U.S. population growth.

c. Make a plot over time of the immigration by decade. Describe the pattern you see here. Can you fit one of the models from this section to data that look like this?

E61. Display 3.116 gives data about passengers on United Airlines flight 815, Chicago-O’Hare to Los Angeles, on October 31, 1997. There were 186 passengers, but the data concern those 33 passengers who had tickets for the Chicago-to-Los Angeles leg only. The variables are

- X: number of days before the flight that the ticket was purchased
- Y: price of the ticket

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>855.97</td>
</tr>
<tr>
<td>7</td>
<td>855.97</td>
</tr>
<tr>
<td>0</td>
<td>1248.51</td>
</tr>
<tr>
<td>20</td>
<td>956.88</td>
</tr>
<tr>
<td>11</td>
<td>125.88</td>
</tr>
<tr>
<td>4</td>
<td>517.05</td>
</tr>
<tr>
<td>77</td>
<td>229.6</td>
</tr>
<tr>
<td>18</td>
<td>165.98</td>
</tr>
<tr>
<td>15</td>
<td>255.91</td>
</tr>
<tr>
<td>14</td>
<td>114.99</td>
</tr>
<tr>
<td>14</td>
<td>114.99</td>
</tr>
<tr>
<td>14</td>
<td>164.44</td>
</tr>
<tr>
<td>14</td>
<td>164.44</td>
</tr>
<tr>
<td>3</td>
<td>137.39</td>
</tr>
<tr>
<td>71</td>
<td>103.46</td>
</tr>
<tr>
<td>71</td>
<td>103.46</td>
</tr>
<tr>
<td>15</td>
<td>168.08</td>
</tr>
</tbody>
</table>

Display 3.116 Number of days before the flight that the ticket was purchased and price of airline ticket. [Source: New York Times, Weekly Review, April 12, 1998.]

Because it is in the airline’s interest to sell tickets early, you might expect Y to be negatively associated with X.

It happens that the first four cases in Display 3.116 are for passengers who flew first class, and those passengers pay more than other
passengers no matter when they purchase their ticket. So you can justify examining the data for the 29 economy-class passengers only.

Finally, one passenger paid $0 because he or she used frequent-flyer miles. You are justified in deleting this value from the data set if the goal is to find a model that relates price to time of purchase.

Can you find a model that relates the cost of the ticket to the number of days in advance that the ticket was purchased? Explain the problems you encounter in doing this.

E62. Different body organs use different amounts of oxygen, even when you take their mass into consideration. For example, the brain uses more oxygen per kilogram of tissue than the lungs do. Scientists are interested in how oxygen consumption is related to the mass of an animal and whether that relationship differs from organ to organ. The data in Display 3.117 show typical body mass, oxygen consumption in brain tissue, and oxygen consumption in lung tissue for a selection of animals. (Oxygen consumption often is measured in milliliters per hour per gram of tissue, but the actual units were not recorded for these data.)

a. As you can see from the table, as total body mass increases, the oxygen consumption in brain tissue tends to go down. Define a function that models this situation. Then find a way to describe the rate of decrease.
b. Repeat part a for lung tissue. How does this relationship differ from that of brain tissue?
c. It is known that the proportion of body mass concentrated in the brain decreases appreciably as the size of the animal increases, whereas the proportion concentrated in the lungs remains relatively constant. One possible theory on oxygen consumption is that the rates of consumption within organ tissue can be explained largely by the relative size of the organ within the body. Is this theory supported by the data? Explain your reasoning.

E63. How is the birthrate of countries related to their economic output? Do richer countries have higher birthrates, perhaps because families can afford more children? Or do poorer countries have higher birthrates, perhaps due to the need for family workers and a lack of education? Display 3.118 shows the birthrates (number of births per thousand population) and the GNP (in thousands of dollars per capita) for a selection of countries from around the world.

a. Construct a scatterplot of these data and comment on the pattern you observe.
b. Fit a statistical model to these data and interpret the slope and intercept of the model in the context of the data.
<table>
<thead>
<tr>
<th>Country</th>
<th>Birthrate (per 1000)</th>
<th>GNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algeria</td>
<td>18.9</td>
<td>1.7</td>
</tr>
<tr>
<td>Argentina</td>
<td>17.7</td>
<td>4.2</td>
</tr>
<tr>
<td>Australia</td>
<td>12.7</td>
<td>19.5</td>
</tr>
<tr>
<td>Brazil</td>
<td>18.1</td>
<td>2.8</td>
</tr>
<tr>
<td>Canada</td>
<td>11.1</td>
<td>22.4</td>
</tr>
<tr>
<td>China</td>
<td>12.8</td>
<td>1.0</td>
</tr>
<tr>
<td>Colombia</td>
<td>22</td>
<td>1.8</td>
</tr>
<tr>
<td>Denmark</td>
<td>12</td>
<td>30.3</td>
</tr>
<tr>
<td>Egypt</td>
<td>24.9</td>
<td>1.5</td>
</tr>
<tr>
<td>France</td>
<td>12.7</td>
<td>22.2</td>
</tr>
<tr>
<td>Germany</td>
<td>8.8</td>
<td>22.7</td>
</tr>
<tr>
<td>India</td>
<td>23.8</td>
<td>0.5</td>
</tr>
<tr>
<td>Indonesia</td>
<td>21.9</td>
<td>0.7</td>
</tr>
<tr>
<td>Israel</td>
<td>18.9</td>
<td>16.0</td>
</tr>
<tr>
<td>Japan</td>
<td>9.6</td>
<td>34.0</td>
</tr>
<tr>
<td>Malaysia</td>
<td>24.2</td>
<td>3.5</td>
</tr>
<tr>
<td>Mexico</td>
<td>22.3</td>
<td>5.9</td>
</tr>
<tr>
<td>Nigeria</td>
<td>39.2</td>
<td>0.3</td>
</tr>
<tr>
<td>Pakistan</td>
<td>32.8</td>
<td>0.4</td>
</tr>
<tr>
<td>Philippines</td>
<td>26.8</td>
<td>1.0</td>
</tr>
<tr>
<td>Russia</td>
<td>9.2</td>
<td>2.1</td>
</tr>
<tr>
<td>South Africa</td>
<td>19.4</td>
<td>2.5</td>
</tr>
<tr>
<td>Spain</td>
<td>10</td>
<td>14.6</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>11.1</td>
<td>25.5</td>
</tr>
<tr>
<td>United States</td>
<td>14.2</td>
<td>35.4</td>
</tr>
</tbody>
</table>


E64. According to the National Center for Health Statistics, the percentage of males who smoke has decreased markedly over the past 40 years, but there still may be some interesting trends to observe. Display 3.119 shows the percentage of males who smoke in selected years for various age groups.

a. Study the trend in the percentage of smokers for the entire male population age 18 and over. The points follow the pattern of exponential decay. How should you modify the percentages before taking their logarithms? Fit the model and interpret the slope.

b. Study the trend in the percentage of smokers for the group age 18 to 24. What model would you use to explain the relationship between the percentage of smokers and the year for this age group? Explain your reasoning. What feature makes this data set more difficult to model than the data set in part a?

c. Study the trend in the percentage of smokers for the group age 65 and over. Does this group show the same kind of trend as seen in the two groups studied in parts a and b? Explain.

Display 3.119 Percentage of males who smoke by age group and year. [Source: National Center for Health Statistics, 2003.]

E65. Is global warming a reality? One measure of global warming is the amount of carbon dioxide (CO₂) in the atmosphere. Display 3.120 gives the annual average carbon dioxide levels (in parts per million) in the atmosphere over Mauna Loa Observatory in Hawaii for the years 1959 through 2003.

a. Plot the data and describe the trend over the years.
b. Fit a straight line to the data and look at the residuals. Describe the pattern you see.

c. Suggest another model that might fit these data well. Fit the model and assess how well it removes the pattern from the residuals.

d. Use the model you like best to describe numerically the growth rate in atmospheric carbon dioxide over Hawaii.

<table>
<thead>
<tr>
<th>Year</th>
<th>CO₂</th>
<th>Year</th>
<th>CO₂</th>
<th>Year</th>
<th>CO₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>1959</td>
<td>316.1</td>
<td>1974</td>
<td>330.4</td>
<td>1989</td>
<td>352.9</td>
</tr>
<tr>
<td>1960</td>
<td>317.0</td>
<td>1975</td>
<td>331.0</td>
<td>1990</td>
<td>354.2</td>
</tr>
<tr>
<td>1961</td>
<td>317.7</td>
<td>1976</td>
<td>332.1</td>
<td>1991</td>
<td>355.6</td>
</tr>
<tr>
<td>1962</td>
<td>318.6</td>
<td>1977</td>
<td>333.6</td>
<td>1992</td>
<td>356.4</td>
</tr>
<tr>
<td>1963</td>
<td>319.1</td>
<td>1978</td>
<td>335.2</td>
<td>1993</td>
<td>357.1</td>
</tr>
<tr>
<td>1964</td>
<td>*</td>
<td>1979</td>
<td>336.5</td>
<td>1994</td>
<td>358.9</td>
</tr>
<tr>
<td>1965</td>
<td>320.4</td>
<td>1980</td>
<td>338.4</td>
<td>1995</td>
<td>361.0</td>
</tr>
<tr>
<td>1966</td>
<td>321.1</td>
<td>1981</td>
<td>339.5</td>
<td>1996</td>
<td>362.6</td>
</tr>
<tr>
<td>1967</td>
<td>322.0</td>
<td>1982</td>
<td>340.8</td>
<td>1997</td>
<td>363.8</td>
</tr>
<tr>
<td>1968</td>
<td>322.8</td>
<td>1983</td>
<td>342.8</td>
<td>1998</td>
<td>366.6</td>
</tr>
<tr>
<td>1969</td>
<td>324.2</td>
<td>1984</td>
<td>344.3</td>
<td>1999</td>
<td>368.3</td>
</tr>
<tr>
<td>1970</td>
<td>325.5</td>
<td>1985</td>
<td>345.7</td>
<td>2000</td>
<td>369.5</td>
</tr>
<tr>
<td>1971</td>
<td>326.5</td>
<td>1986</td>
<td>346.9</td>
<td>2001</td>
<td>371.0</td>
</tr>
<tr>
<td>1972</td>
<td>327.6</td>
<td>1987</td>
<td>348.6</td>
<td>2002</td>
<td>373.1</td>
</tr>
<tr>
<td>1973</td>
<td>329.8</td>
<td>1988</td>
<td>351.2</td>
<td>2003</td>
<td>375.6</td>
</tr>
</tbody>
</table>

Display 3.120 Carbon dioxide in the atmosphere. [Source: Mauna Loa Observatory.]

E66. How does the average SAT math score for students in a state relate to the percentage of students taking the exam? Display 3.121 shows the average SAT math score for each state in 2005, along with the percentage of high school seniors taking the exam. Find a model that seems like a good predictor of average SAT math scores based on knowledge of the percentage of seniors taking the exam.

<table>
<thead>
<tr>
<th>State</th>
<th>Math SAT Score</th>
<th>% Seniors Taking Exam</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alaska</td>
<td>519</td>
<td>53</td>
</tr>
<tr>
<td>Alabama</td>
<td>559</td>
<td>10</td>
</tr>
<tr>
<td>Arkansas</td>
<td>552</td>
<td>6</td>
</tr>
<tr>
<td>Arizona</td>
<td>530</td>
<td>33</td>
</tr>
<tr>
<td>California</td>
<td>522</td>
<td>50</td>
</tr>
<tr>
<td>Colorado</td>
<td>560</td>
<td>26</td>
</tr>
<tr>
<td>Connecticut</td>
<td>517</td>
<td>86</td>
</tr>
<tr>
<td>Delaware</td>
<td>502</td>
<td>74</td>
</tr>
<tr>
<td>Florida</td>
<td>498</td>
<td>65</td>
</tr>
<tr>
<td>Georgia</td>
<td>496</td>
<td>75</td>
</tr>
<tr>
<td>Hawaii</td>
<td>516</td>
<td>61</td>
</tr>
<tr>
<td>Iowa</td>
<td>608</td>
<td>5</td>
</tr>
<tr>
<td>Idaho</td>
<td>542</td>
<td>21</td>
</tr>
<tr>
<td>Illinois</td>
<td>606</td>
<td>10</td>
</tr>
<tr>
<td>Indiana</td>
<td>508</td>
<td>66</td>
</tr>
<tr>
<td>Kansas</td>
<td>588</td>
<td>9</td>
</tr>
<tr>
<td>Kentucky</td>
<td>559</td>
<td>12</td>
</tr>
<tr>
<td>Louisiana</td>
<td>562</td>
<td>8</td>
</tr>
<tr>
<td>Massachusetts</td>
<td>527</td>
<td>86</td>
</tr>
<tr>
<td>Maryland</td>
<td>515</td>
<td>71</td>
</tr>
<tr>
<td>Maine</td>
<td>505</td>
<td>75</td>
</tr>
<tr>
<td>Michigan</td>
<td>579</td>
<td>10</td>
</tr>
<tr>
<td>Mississippi</td>
<td>554</td>
<td>4</td>
</tr>
<tr>
<td>Minnesota</td>
<td>597</td>
<td>11</td>
</tr>
<tr>
<td>Missouri</td>
<td>588</td>
<td>7</td>
</tr>
<tr>
<td>Montana</td>
<td>540</td>
<td>31</td>
</tr>
<tr>
<td>North Carolina</td>
<td>511</td>
<td>74</td>
</tr>
<tr>
<td>North Dakota</td>
<td>605</td>
<td>4</td>
</tr>
<tr>
<td>Nebraska</td>
<td>579</td>
<td>8</td>
</tr>
<tr>
<td>New Hampshire</td>
<td>525</td>
<td>81</td>
</tr>
<tr>
<td>New Jersey</td>
<td>517</td>
<td>86</td>
</tr>
<tr>
<td>New Mexico</td>
<td>547</td>
<td>13</td>
</tr>
<tr>
<td>Nevada</td>
<td>513</td>
<td>39</td>
</tr>
<tr>
<td>New York</td>
<td>511</td>
<td>92</td>
</tr>
<tr>
<td>Ohio</td>
<td>543</td>
<td>29</td>
</tr>
<tr>
<td>Oklahoma</td>
<td>563</td>
<td>7</td>
</tr>
<tr>
<td>Oregon</td>
<td>528</td>
<td>59</td>
</tr>
<tr>
<td>Pennsylvania</td>
<td>503</td>
<td>75</td>
</tr>
<tr>
<td>Rhode Island</td>
<td>505</td>
<td>72</td>
</tr>
<tr>
<td>South Carolina</td>
<td>499</td>
<td>64</td>
</tr>
<tr>
<td>South Dakota</td>
<td>589</td>
<td>5</td>
</tr>
<tr>
<td>Tennessee</td>
<td>563</td>
<td>16</td>
</tr>
<tr>
<td>Texas</td>
<td>502</td>
<td>54</td>
</tr>
<tr>
<td>Utah</td>
<td>557</td>
<td>7</td>
</tr>
<tr>
<td>Virginia</td>
<td>514</td>
<td>73</td>
</tr>
<tr>
<td>Vermont</td>
<td>517</td>
<td>67</td>
</tr>
<tr>
<td>Washington</td>
<td>534</td>
<td>55</td>
</tr>
<tr>
<td>Wisconsin</td>
<td>599</td>
<td>6</td>
</tr>
<tr>
<td>West Virginia</td>
<td>511</td>
<td>20</td>
</tr>
<tr>
<td>Wyoming</td>
<td>543</td>
<td>12</td>
</tr>
</tbody>
</table>

Display 3.121 Average SAT math scores by state. [Source: College Board, www.collegeboard.com.]
Chapter Summary

In Chapter 2, you worked with univariate data. In Chapter 3, you learned how to uncover information for bivariate (two-variable) data, using plots and numerical summaries of center and spread. In Chapter 2, the basic plot was the histogram. For histograms of distributions that are approximately normal, the fundamental measures of center and spread are the mean and the standard deviation. For bivariate data, the basic plot is the scatterplot. For scatterplots that have an elliptical shape, the fundamental summary measures are the regression line (which you can think of as the measure of center) and the correlation (which you can think of as the measure of spread).

For now, correlation and regression merely describe your data set. In Chapter 11, you will learn to use numerical summaries computed from a sample to make inferences about a larger population. Using diagnostic tools such as residual plots and finding transformations that re-express a curved relationship as a linear one will come in especially handy because you won’t be able to make valid inferences unless the points form an elliptical cloud.

Review Exercises

E67. Leonardo’s rules. A class of 15 students recorded the measurements in Display 3.122 for Activity 3.3a.

<table>
<thead>
<tr>
<th>Student</th>
<th>Height</th>
<th>Arm Span</th>
<th>Kneeling Height</th>
<th>Hand Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>170.5</td>
<td>168.0</td>
<td>126.0</td>
<td>18.0</td>
</tr>
<tr>
<td>2</td>
<td>170.0</td>
<td>172.0</td>
<td>129.5</td>
<td>18.0</td>
</tr>
<tr>
<td>3</td>
<td>107.0</td>
<td>101.0</td>
<td>79.5</td>
<td>10.0</td>
</tr>
<tr>
<td>4</td>
<td>159.0</td>
<td>161.0</td>
<td>116.0</td>
<td>16.0</td>
</tr>
<tr>
<td>5</td>
<td>166.0</td>
<td>166.0</td>
<td>122.0</td>
<td>18.0</td>
</tr>
<tr>
<td>6</td>
<td>175.0</td>
<td>174.0</td>
<td>125.0</td>
<td>19.5</td>
</tr>
<tr>
<td>7</td>
<td>158.0</td>
<td>153.5</td>
<td>116.0</td>
<td>16.0</td>
</tr>
<tr>
<td>8</td>
<td>95.5</td>
<td>95.0</td>
<td>71.5</td>
<td>10.0</td>
</tr>
<tr>
<td>9</td>
<td>132.5</td>
<td>129.0</td>
<td>95.0</td>
<td>11.5</td>
</tr>
<tr>
<td>10</td>
<td>165.0</td>
<td>169.0</td>
<td>124.0</td>
<td>17.0</td>
</tr>
<tr>
<td>11</td>
<td>179.0</td>
<td>175.0</td>
<td>131.0</td>
<td>20.0</td>
</tr>
<tr>
<td>12</td>
<td>149.0</td>
<td>154.0</td>
<td>109.5</td>
<td>15.5</td>
</tr>
<tr>
<td>13</td>
<td>143.0</td>
<td>142.0</td>
<td>111.5</td>
<td>16.0</td>
</tr>
<tr>
<td>14</td>
<td>158.0</td>
<td>156.5</td>
<td>119.0</td>
<td>17.5</td>
</tr>
<tr>
<td>15</td>
<td>161.0</td>
<td>164.0</td>
<td>121.0</td>
<td>16.5</td>
</tr>
</tbody>
</table>

Display 3.122 Sample measurements, in centimeters, for Activity 3.3a.

a. Construct scatterplots and fit least squares lines for each of Leonardo’s rules in Activity 3.3a. Do the rules appear to hold?
b. Interpret the slopes of your regression lines.
c. If appropriate, find the value of $r$ for each of the three relationships. Which correlation is strongest? Which is weakest?

E68. Space Shuttle Challenger. On January 28, 1986, because two O-rings did not seal properly, Space Shuttle Challenger exploded and seven people died. The temperature predicted for the morning of the flight was between 26°F and 29°F. The engineers were concerned that the cold temperatures would cause the rubber O-rings to malfunction. On seven previous flights at least one of the twelve O-rings had shown some distress. The NASA officials and engineers who decided not to delay the flight had available to them data like those on the scatterplot in Display 3.123 before they made that decision.
**Display 3.123** Flights when at least one O-ring showed some distress.

a. Why did it seem reasonable to launch despite the low temperature?

b. Display 3.123 contains information only about flights that had O-ring failures. Data for all flights were available on a table like the one in Display 3.124. Add the missing points to a copy of the scatterplot in Display 3.123. How do these data affect any trend in the scatterplot? Would you have recommended launching the space shuttle if you had seen the complete plot? Why or why not?


<table>
<thead>
<tr>
<th>Temperature (°F)</th>
<th>Number of O-Rings with Some Distress</th>
</tr>
</thead>
<tbody>
<tr>
<td>53</td>
<td>3</td>
</tr>
<tr>
<td>57</td>
<td>1</td>
</tr>
<tr>
<td>58</td>
<td>1</td>
</tr>
<tr>
<td>63</td>
<td>1</td>
</tr>
<tr>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>75</td>
<td>2</td>
</tr>
<tr>
<td>66</td>
<td>0</td>
</tr>
<tr>
<td>67</td>
<td>0</td>
</tr>
<tr>
<td>67</td>
<td>0</td>
</tr>
<tr>
<td>68</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Temperature (°F)</th>
<th>Number of O-Rings with Some Distress</th>
</tr>
</thead>
<tbody>
<tr>
<td>69</td>
<td>0</td>
</tr>
<tr>
<td>70</td>
<td>0</td>
</tr>
<tr>
<td>70</td>
<td>0</td>
</tr>
<tr>
<td>72</td>
<td>0</td>
</tr>
<tr>
<td>73</td>
<td>0</td>
</tr>
<tr>
<td>75</td>
<td>0</td>
</tr>
<tr>
<td>76</td>
<td>0</td>
</tr>
<tr>
<td>76</td>
<td>0</td>
</tr>
<tr>
<td>78</td>
<td>0</td>
</tr>
<tr>
<td>78</td>
<td>0</td>
</tr>
<tr>
<td>79</td>
<td>0</td>
</tr>
<tr>
<td>80</td>
<td>0</td>
</tr>
<tr>
<td>81</td>
<td>0</td>
</tr>
</tbody>
</table>

**Display 3.125** Data for exam scores in a statistics class, with scatterplot and residual plot.

a. Is there a point that is more influential than the other points on the slope of the regression line? How can you tell from the scatterplot? From the residual plot?

b. How will the slope change if the scores for this one influential point are removed from the data set? How will the correlation change? Calculate the slope and correlation for the revised data to check your estimate.

**E69. Exam scores.** Students’ scores on two exams in a statistics course are given in Display 3.125 along with a scatterplot with regression line and a residual plot. The regression equation is \( \text{Exam 2} = 51.0 + 0.430(\text{Exam 1}) \), and the correlation, \( r \), is 0.756.
c. Construct a residual plot of the revised data. Does a linear model fit the data well?

d. Refer to the scatterplot of Exam 2 versus Exam 1 in Display 3.125. Does this plot illustrate regression to the mean? Explain your reasoning.

E70. Suppose you have the Exam 1 and Exam 2 scores of all students enrolled in U.S. History.

a. The slope of the regression line for predicting the scores on Exam 2 from the scores on Exam 1 is 0.51. The standard deviation for Exam 1 scores is 11.6, and the standard deviation for Exam 2 scores is 7.0. Use only this information to find the correlation coefficient for these scores.

b. Suppose you know, in addition, that the means are 82.3 for Exam 1 and 87.8 for Exam 2. Find the equation of the least squares line for predicting Exam 2 scores from Exam 1 scores.

E71. You are given a list of six values, −1.5, −0.5, 0, 0, 0.5, and 1.5, for x and the same list of six values for y. Note that the list has mean 0 and standard deviation 1.

a. Match each x-value with a y-value so that the resulting six pairs (x, y) have correlation 1.

b. Match the x- and y-values again so that the points have the largest possible correlation less than 1.

c. Match the values again, this time to get a correlation as close to 0 as possible.

d. Match the values a fourth time to get a correlation of −1.

E72. Display 3.126 lists the values of six variables, with a “scatterplot matrix” showing all 30 possible scatterplots for these variables. For example, the first scatterplot in the first row has variable B on the x-axis and variable A on the y-axis. The first scatterplot in the second row has variable A on the x-axis and variable B on the y-axis.

a. For five pairs of variables the correlation is exactly 0, and for one other pair it is 0.02, or almost 0. Identify these six pairs of variables. What do they have in common?

b. At the other extreme, one pair of variables has correlation 0.87; the next highest correlation is 0.58, and the third highest is 0.45. Identify these three pairs, and put them in order from strongest to weakest correlation.

c. Of the remaining six pairs, four have correlations of about 0.25 (give or take a little) and two have correlations of about

### Display 3.126

Data table for 6 variables and a "scatterplot matrix" of all 30 possible scatterplots for the variables.
0.1 (give or take a little). Which four pairs have correlations around 0.25?

d. Choose several scatterplots that you think best illustrate the phrase “the correlation measures direction and strength but not shape,” and use them to show what you mean.

E73. Decide whether each statement is true or false, and then explain your decision.

a. The correlation is to bivariate data what the standard deviation is to univariate data.
b. The correlation measures direction and strength but not shape.
c. If the correlation is near 0, knowing the value of one variable gives you a narrow interval of likely values for the other variable.
d. No matter what data set you look at, the correlation coefficient, $r$, and least squares slope, $b_1$, will always have the same sign.

E74. Look at the scatterplot of average SAT I math scores versus the percentage of students taking the exam in Display 3.7 on page 112.

a. Estimate the correlation.
b. What possibly important features of the plot are lost if you give only the correlation and the equation of the least squares line?
c. Sketch what you think the residual plot would look like if you fitted one line to all the points.

E75. The correlation between in-state tuition and out-of-state tuition, measured in dollars, for a sample of public universities is 0.80.

a. Rewrite the sentence above so that someone who does not know statistics can understand it.
b. Does the correlation change if you convert tuition costs to thousands of dollars and recompute the slope? Does it change if you take logarithms of the tuition costs and recompute the slope?

c. Does the slope of the least squares line change if you convert tuition costs to thousands of dollars and recompute the slope? Does it change if you take logarithms of the tuition costs and recompute the slope?

E76. Display 3.127 shows a scatterplot divided into quadrants by vertical and horizontal lines that pass through the point of averages, $(\bar{x}, \bar{y})$.

a. For each of the four quadrants, give the sign of $z_x$ (the standardized value of $x$), $z_y$ (the standardized value of $y$), and their product $z_x \cdot z_y$.
b. Which point(s) make the smallest contribution to the correlation? Explain why the contributions are small.

![Display 3.127](image)

**Display 3.127** A scatterplot divided into quadrants by lines passing through the means.

E77. Rank these summaries for three sets of bivariate data by the strength of the relationship, from weakest to strongest.

A. $\hat{y} = 90 + 100x$  
   
   $s_x = 5$  
   
   $s_y = 1000$

B. $\hat{y} = \frac{x}{3} - 12$  
   
   $s_x = 0.9$  
   
   $s_y = 1$

C. $\hat{y} = 1.05 + 0.01x$  
   
   $s_x = 0.05$  
   
   $s_y = 0.002$
E78. There’s an extremely strong relationship between the price of books online and the price at your local bookstore.

a. Does this mean the prices are almost the same?

b. Explain why it is wrong to say that the prices online “cause” the prices at your local bookstore. Why is the relationship so strong if neither set of prices causes the other?

e79. Describe a set of cases and two variables for which you would expect to see regression toward the mean.

E80. Life spans. In Chapter 2, you looked at the characteristics of mammals (given in Display 2.24 on page 43) one at a time. Now you can look at the relationship between two variables. For example, is longevity associated with gestation period? The variables are average longevity in years, maximum longevity in years, gestation period in days, and speed in miles per hour.

a. Construct a scatterplot of gestation period versus maximum longevity. Describe what you see, including an estimate of the correlation.

b. Repeat part a, with average longevity in place of maximum longevity. Does the average longevity or the maximum longevity give a better prediction of the gestation period?

c. Does speed appear to be associated with average longevity?

E81. Spending for police. The data in Display 3.128 give the number of police officers, the total expenditures for police officers, the population, and the violent crime rate for a sample of states in 2000.

a. Explore and summarize the relationship between the number of police officers and total expenditures for police.

b. Explore and summarize the relationship between the population of the states and the number of police officers they employ.

c. Is the number of police officers strongly related to the rate of violent crime in these states? Explain. Find a transformation that straightens these data. Check the linearity of your transformed data with a residual plot.

<table>
<thead>
<tr>
<th>State</th>
<th>Number of Police Officers (in thousands)</th>
<th>Expenditures for Police (in millions of dollars)</th>
<th>Population (in millions)</th>
<th>Violent Crime Rate (number per 100,000 of state population)</th>
</tr>
</thead>
<tbody>
<tr>
<td>California</td>
<td>96.9</td>
<td>7653</td>
<td>34.0</td>
<td>622</td>
</tr>
<tr>
<td>Colorado</td>
<td>12.0</td>
<td>753</td>
<td>4.3</td>
<td>334</td>
</tr>
<tr>
<td>Florida</td>
<td>55.2</td>
<td>3371</td>
<td>16.0</td>
<td>812</td>
</tr>
<tr>
<td>Illinois</td>
<td>44.1</td>
<td>2718</td>
<td>12.4</td>
<td>657</td>
</tr>
<tr>
<td>Iowa</td>
<td>7.3</td>
<td>346</td>
<td>2.9</td>
<td>266</td>
</tr>
<tr>
<td>Louisiana</td>
<td>16.1</td>
<td>635</td>
<td>4.5</td>
<td>681</td>
</tr>
<tr>
<td>Maine</td>
<td>3.1</td>
<td>118</td>
<td>1.3</td>
<td>110</td>
</tr>
<tr>
<td>Mississippi</td>
<td>8.6</td>
<td>337</td>
<td>2.8</td>
<td>361</td>
</tr>
<tr>
<td>New Jersey</td>
<td>33.4</td>
<td>1829</td>
<td>8.4</td>
<td>384</td>
</tr>
<tr>
<td>Tennessee</td>
<td>18.1</td>
<td>828</td>
<td>5.7</td>
<td>707</td>
</tr>
<tr>
<td>Texas</td>
<td>58.9</td>
<td>2866</td>
<td>20.9</td>
<td>545</td>
</tr>
<tr>
<td>Virginia</td>
<td>18.8</td>
<td>965</td>
<td>7.1</td>
<td>282</td>
</tr>
<tr>
<td>Washington</td>
<td>14.1</td>
<td>854</td>
<td>5.9</td>
<td>370</td>
</tr>
</tbody>
</table>

E82. *House prices.* Display 3.129 gives the selling prices for all houses sold in a Florida community in one month.

a. Construct a model to predict the selling price from the area, transforming any variables, if necessary. Would you use the same model for both new and used houses?

b. Are there any influential observations that have a serious effect on the model? If so, what would happen to the slope of the prediction equation and the correlation if you removed this (or these) point(s) from the analysis?

c. Predict the selling price of an old house measuring 1000 sq ft. Do the same for an old house measuring 2000 sq ft. Which prediction do you feel more confident about? Explain.

d. Explain the effect of the number of bathrooms on the selling price of the houses. Is it appropriate to fit a regression model to price as a function of the number of bathrooms and interpret the results in the usual way? Why or why not?

<table>
<thead>
<tr>
<th>House</th>
<th>Price ($ thousands)</th>
<th>Area (thousands of sq ft)</th>
<th>Number of Bedrooms</th>
<th>Number of Bathrooms</th>
<th>New (1), Old (0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>48.5</td>
<td>1.10</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>55.0</td>
<td>1.01</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>68.0</td>
<td>1.45</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>137.0</td>
<td>2.40</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>309.4</td>
<td>3.30</td>
<td>4</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>17.5</td>
<td>0.40</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>19.6</td>
<td>1.28</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>24.5</td>
<td>0.74</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>34.8</td>
<td>0.78</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>32.0</td>
<td>0.97</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>28.0</td>
<td>0.84</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>49.9</td>
<td>1.08</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>59.9</td>
<td>0.99</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>61.5</td>
<td>1.01</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>60.0</td>
<td>1.34</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>65.9</td>
<td>1.22</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>67.9</td>
<td>1.28</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>68.9</td>
<td>1.29</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>69.9</td>
<td>1.52</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>70.5</td>
<td>1.25</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>72.9</td>
<td>1.28</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>22</td>
<td>72.5</td>
<td>1.28</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>23</td>
<td>72.0</td>
<td>1.36</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>24</td>
<td>71.0</td>
<td>1.20</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>25</td>
<td>76.0</td>
<td>1.46</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>26</td>
<td>72.9</td>
<td>1.56</td>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>27</td>
<td>73.0</td>
<td>1.22</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>28</td>
<td>70.0</td>
<td>1.40</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>29</td>
<td>76.0</td>
<td>1.15</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>30</td>
<td>69.0</td>
<td>1.74</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>31</td>
<td>75.5</td>
<td>1.62</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>32</td>
<td>76.0</td>
<td>1.66</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>33</td>
<td>81.8</td>
<td>1.33</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>34</td>
<td>84.5</td>
<td>1.34</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>35</td>
<td>83.5</td>
<td>1.40</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Display 3.129 (continued)
<table>
<thead>
<tr>
<th>House</th>
<th>Price ($ thousands)</th>
<th>Area (thousands of sq ft)</th>
<th>Number of Bedrooms</th>
<th>Number of Bathrooms</th>
<th>New (1), Old (0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>36</td>
<td>86.0</td>
<td>1.15</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>37</td>
<td>86.9</td>
<td>1.58</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>38</td>
<td>86.9</td>
<td>1.58</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>39</td>
<td>86.9</td>
<td>1.58</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>40</td>
<td>87.9</td>
<td>1.71</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>41</td>
<td>88.1</td>
<td>2.10</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>42</td>
<td>85.9</td>
<td>1.27</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>43</td>
<td>89.5</td>
<td>1.34</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>44</td>
<td>87.4</td>
<td>1.25</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>45</td>
<td>87.9</td>
<td>1.68</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>46</td>
<td>88.0</td>
<td>1.55</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>47</td>
<td>90.0</td>
<td>1.55</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>48</td>
<td>96.0</td>
<td>1.36</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>49</td>
<td>99.9</td>
<td>1.51</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>50</td>
<td>95.5</td>
<td>1.54</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>51</td>
<td>98.5</td>
<td>1.51</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>52</td>
<td>100.1</td>
<td>1.85</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>53</td>
<td>99.9</td>
<td>1.62</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>54</td>
<td>101.9</td>
<td>1.40</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>55</td>
<td>101.9</td>
<td>1.92</td>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>56</td>
<td>102.3</td>
<td>1.42</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>57</td>
<td>110.8</td>
<td>1.56</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>58</td>
<td>105.0</td>
<td>1.43</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>59</td>
<td>97.9</td>
<td>2.00</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>60</td>
<td>106.3</td>
<td>1.45</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>61</td>
<td>106.5</td>
<td>1.65</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>62</td>
<td>116.0</td>
<td>1.72</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>63</td>
<td>108.0</td>
<td>1.79</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>64</td>
<td>107.5</td>
<td>1.85</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>65</td>
<td>109.9</td>
<td>2.06</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>66</td>
<td>110.0</td>
<td>1.76</td>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>67</td>
<td>120.0</td>
<td>1.62</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>68</td>
<td>115.0</td>
<td>1.80</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>69</td>
<td>113.4</td>
<td>1.98</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>70</td>
<td>114.9</td>
<td>1.57</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>71</td>
<td>115.0</td>
<td>2.19</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>72</td>
<td>115.0</td>
<td>2.07</td>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>73</td>
<td>117.9</td>
<td>1.99</td>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>74</td>
<td>110.0</td>
<td>1.55</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>75</td>
<td>115.0</td>
<td>1.67</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>76</td>
<td>124.0</td>
<td>2.40</td>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>77</td>
<td>129.9</td>
<td>1.79</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>78</td>
<td>124.0</td>
<td>1.89</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>79</td>
<td>128.0</td>
<td>1.88</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>80</td>
<td>132.4</td>
<td>2.00</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>81</td>
<td>139.3</td>
<td>2.05</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>82</td>
<td>139.3</td>
<td>2.00</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>83</td>
<td>139.7</td>
<td>2.03</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>84</td>
<td>142.0</td>
<td>2.12</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>85</td>
<td>141.3</td>
<td>2.08</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Display 3.129 (continued)
Examine the association between per-pupil expenditure and average teacher salary, with the goal of predicting per-pupil expenditure. Is this a cause-and-effect relationship?

b. Analyze the effect of average teacher salary on per-capita expenditure (spending on public schools divided by the number of people in the state). Compare the association to the association in part a. Are the relative sizes of the correlations about what you would expect?

c. Are any variables good predictors of the percentage of dropouts? Explain your reasoning.

<table>
<thead>
<tr>
<th>State</th>
<th>ExpPP</th>
<th>ExpPC</th>
<th>TeaSal</th>
<th>%Dropout</th>
<th>Enroll</th>
<th>Teachers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iowa</td>
<td>8064</td>
<td>1335</td>
<td>39.1</td>
<td>2.7</td>
<td>486</td>
<td>34.6</td>
</tr>
<tr>
<td>Kansas</td>
<td>8568</td>
<td>1485</td>
<td>37.8</td>
<td>3.2</td>
<td>470</td>
<td>32.6</td>
</tr>
<tr>
<td>Kentucky</td>
<td>7619</td>
<td>1218</td>
<td>39</td>
<td>4.6</td>
<td>654</td>
<td>38.8</td>
</tr>
<tr>
<td>Louisiana</td>
<td>8081</td>
<td>1320</td>
<td>37.2</td>
<td>8.3</td>
<td>731</td>
<td>50.3</td>
</tr>
<tr>
<td>Maine</td>
<td>10170</td>
<td>1618</td>
<td>38.5</td>
<td>3.1</td>
<td>206</td>
<td>16</td>
</tr>
<tr>
<td>Maryland</td>
<td>9242</td>
<td>1460</td>
<td>49.7</td>
<td>4.1</td>
<td>861</td>
<td>55.5</td>
</tr>
<tr>
<td>Massachusetts</td>
<td>11403</td>
<td>1728</td>
<td>50.8</td>
<td>3.4</td>
<td>973</td>
<td>71.9</td>
</tr>
<tr>
<td>Michigan</td>
<td>10377</td>
<td>1788</td>
<td>53.6</td>
<td>n/a</td>
<td>1731</td>
<td>92.2</td>
</tr>
<tr>
<td>Minnesota</td>
<td>10888</td>
<td>1844</td>
<td>44.7</td>
<td>4</td>
<td>851</td>
<td>52.6</td>
</tr>
<tr>
<td>Mississippi</td>
<td>6555</td>
<td>1129</td>
<td>34.6</td>
<td>4.6</td>
<td>494</td>
<td>30.6</td>
</tr>
<tr>
<td>Missouri</td>
<td>8188</td>
<td>1314</td>
<td>37.7</td>
<td>4.2</td>
<td>910</td>
<td>66.2</td>
</tr>
<tr>
<td>Montana</td>
<td>7717</td>
<td>1288</td>
<td>35.8</td>
<td>4.2</td>
<td>152</td>
<td>10.4</td>
</tr>
<tr>
<td>Nebraska</td>
<td>8067</td>
<td>1331</td>
<td>37.9</td>
<td>4</td>
<td>285</td>
<td>20.7</td>
</tr>
<tr>
<td>Nevada</td>
<td>8793</td>
<td>1448</td>
<td>41.8</td>
<td>5.2</td>
<td>357</td>
<td>19.5</td>
</tr>
<tr>
<td>New Hampshire</td>
<td>9193</td>
<td>1493</td>
<td>41.9</td>
<td>5.4</td>
<td>207</td>
<td>15</td>
</tr>
<tr>
<td>New Jersey</td>
<td>11822</td>
<td>1850</td>
<td>54.2</td>
<td>2.8</td>
<td>1342</td>
<td>105.4</td>
</tr>
<tr>
<td>New Mexico</td>
<td>8722</td>
<td>1507</td>
<td>37</td>
<td>5.3</td>
<td>320</td>
<td>21</td>
</tr>
<tr>
<td>New York</td>
<td>12918</td>
<td>1939</td>
<td>53</td>
<td>3.8</td>
<td>2872</td>
<td>226</td>
</tr>
<tr>
<td>North Carolina</td>
<td>7925</td>
<td>1255</td>
<td>43.1</td>
<td>6.3</td>
<td>1315</td>
<td>86.1</td>
</tr>
<tr>
<td>North Dakota</td>
<td>7338</td>
<td>1261</td>
<td>33.9</td>
<td>2.2</td>
<td>106</td>
<td>7.7</td>
</tr>
<tr>
<td>Ohio</td>
<td>10028</td>
<td>1609</td>
<td>45.5</td>
<td>3.9</td>
<td>1831</td>
<td>122.2</td>
</tr>
<tr>
<td>Oklahoma</td>
<td>6733</td>
<td>1200</td>
<td>34.9</td>
<td>5.2</td>
<td>622</td>
<td>40.6</td>
</tr>
<tr>
<td>Oregon</td>
<td>8490</td>
<td>1329</td>
<td>47.6</td>
<td>5.3</td>
<td>551</td>
<td>27.1</td>
</tr>
<tr>
<td>Pennsylvania</td>
<td>8735</td>
<td>1291</td>
<td>51.4</td>
<td>3.6</td>
<td>1822</td>
<td>118.3</td>
</tr>
<tr>
<td>Rhode Island</td>
<td>10348</td>
<td>1530</td>
<td>51.1</td>
<td>5</td>
<td>158</td>
<td>13.4</td>
</tr>
<tr>
<td>South Carolina</td>
<td>9009</td>
<td>1517</td>
<td>40.4</td>
<td>3.3</td>
<td>691</td>
<td>44.5</td>
</tr>
<tr>
<td>South Dakota</td>
<td>8117</td>
<td>1367</td>
<td>32.4</td>
<td>3.9</td>
<td>128</td>
<td>9</td>
</tr>
<tr>
<td>Tennessee</td>
<td>6284</td>
<td>1004</td>
<td>39.7</td>
<td>4.3</td>
<td>925</td>
<td>58.3</td>
</tr>
<tr>
<td>Texas</td>
<td>8780</td>
<td>1682</td>
<td>40</td>
<td>4.2</td>
<td>4163</td>
<td>288.4</td>
</tr>
<tr>
<td>Utah</td>
<td>6064</td>
<td>1268</td>
<td>38.3</td>
<td>3.7</td>
<td>485</td>
<td>21.6</td>
</tr>
<tr>
<td>Vermont</td>
<td>11198</td>
<td>1834</td>
<td>41.5</td>
<td>4.7</td>
<td>101</td>
<td>8.8</td>
</tr>
<tr>
<td>Virginia</td>
<td>7165</td>
<td>1143</td>
<td>43.2</td>
<td>3.5</td>
<td>1163</td>
<td>93.1</td>
</tr>
<tr>
<td>Washington</td>
<td>9102</td>
<td>1514</td>
<td>45</td>
<td>4</td>
<td>1009</td>
<td>53.1</td>
</tr>
<tr>
<td>West Virginia</td>
<td>9859</td>
<td>1546</td>
<td>38.5</td>
<td>4.2</td>
<td>283</td>
<td>19.9</td>
</tr>
<tr>
<td>Wisconsin</td>
<td>10631</td>
<td>1718</td>
<td>42.8</td>
<td>2.3</td>
<td>879</td>
<td>60</td>
</tr>
<tr>
<td>Wyoming</td>
<td>10943</td>
<td>1931</td>
<td>38.8</td>
<td>6.4</td>
<td>88</td>
<td>6.5</td>
</tr>
</tbody>
</table>

AP1. This scatterplot shows the age in years of the oldest and youngest child in 116 households with two or more children age 18 or younger living with their parents. (Some points have been moved slightly to show that multiple households are at each coordinate.) Which of the following is not a reasonable interpretation of this scatterplot?

- A. There aren’t any points in the upper-left region because older children tend to move out to go to college or to get married.
- B. The older the oldest child in a household, the older the youngest child tends to be.
- C. There are no households represented here in which the only children are twins.
- D. The variability in the age of the youngest child in these households tends to increase with the age of the oldest child.
- E. Few households have a range of more than 12 years in the ages of all of the children in the household.

AP2. In a study of 190 nations, the least squares line for the relationship between birthrate (per thousand per year) and female literacy rate (in percent) is \( \text{birthrate} = -0.38 \cdot \text{literacy} + 53.5 \), with \( r = 0.8 \). Uganda has a birthrate of 47 and a female literacy rate of 60. What is the residual for Uganda?

- A. -17.1
- B. 29.3
- C. 16.3
- D. 64.1
- E. 69.8

AP3. In a linear regression of the heights of a group of trees versus their circumferences, the pattern of residuals is U-shaped. Which of the following must be true?

I. A nonlinear regression would be a better model.
II. For trees near the middle of the range of tree circumferences studied, the predicted tree height tends to be too tall.
III. \( r \) will be close to 0.

- A. II only
- B. III only
- C. I and II
- D. I and III
- E. I, II, and III.

AP4. A recent study models the relationship between the number of teachers at a high school and the number of sick days these teachers take in a year. This scatterplot shows data for all high schools in a county during one year.

Which of the following is not a reasonable way to proceed with the analysis?

- A. Remove the four outliers permanently from the data set.
- B. Run the regression again without the four outliers to see how much the slope and correlation change.
- C. Verify the values for the four outliers to make sure they are correct.
- D. Try to find a transformation that makes the cloud of points more elliptical.
- E. Make a residual plot in order to judge the linearity of the points.
AP5. Upon checking out of a large hospital, 2000 patients rated their satisfaction with their stay on a scale of 0–10, with 10 indicating complete satisfaction. The relationship between the satisfaction rating and the patient’s length of stay (in days) was analyzed using linear regression. Here is part of the computer printout for this regression.

The regression equation is

\[ \text{Satisfaction} = 4.10 + 0.231 \times \text{Stay} \]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>4.1018</td>
<td>0.3050</td>
<td>13.45</td>
<td>0.000</td>
</tr>
<tr>
<td>Stay</td>
<td>0.23110</td>
<td>0.07979</td>
<td>2.90</td>
<td>0.004</td>
</tr>
</tbody>
</table>

\[ S = 1.50287 \quad R^2 = 5.2\% \quad R^2 (adj) = 4.8\% \]

Which is a correct interpretation of this regression?

A. Patients who stay longer in the hospital tend to be more satisfied than patients with shorter stays, although this relationship is weak.
B. The correlation is ±0.228. However, the sign on the correlation cannot be determined.
C. The value of \( R^2 \) indicates that the relationship between satisfaction rating and length of stay is weak but linear.
D. The \( y \)-intercept of the regression line indicates that no patients rated their satisfaction less than 4.
E. The slope of the regression line indicates that as a patient stays longer in the hospital, his or her satisfaction tends to increase day by day.

AP6. The least squares equation to estimate the population of the fictional country of Barbaria is \( \log_{10}(\text{population}) = 0.01t + 7 \), where \( t \) is the number of years since 1950. Which of the following is closest to the predicted population of Barbaria for the year 2000?

A. 7.5  
B. 27  
C. 7,500,000  
D. 31,600,000  
E. 1,000,000,000,000,000,000,000,000,000

AP7. The Barbarian Aptitude Test (BAT) gives each Barbarian two scores, one for pillaging and one for burning. The scores range from a low of 0 to a high of 50. The least squares equation for a large group of Barbarians who took the BAT is \( \text{burning} = 0.3 \times \text{pillaging} + 19 \), with \( r = 0.6 \). Which is the best interpretation of the slope of this line?

A. A Barbarian who studies harder and improves her pillaging score by 1 point on the next BAT will tend to increase her burning score by about 0.3 point as well.
B. Barbarians tend to score about 0.3 point higher on burning than on pillaging.
C. Barbarians score about 30% as many points on burning as on pillaging.
D. The burning score is highly correlated with the pillaging score.
E. A Barbarian who earned 1 more point on pillaging than another Barbarian tended to earn only 0.3 point more on burning.

AP8. A least squares regression analysis using a rating of each Barbarian’s personal cleanliness as the explanatory variable and the number of raids he or she has carried out as the response variable found a positive relationship with \( R^2 = 0.81 \). Which is not a correct interpretation of this information?

A. The correlation between personal cleanliness and the number of raids is 0.9.
B. There is a strong relationship between personal cleanliness and number of raids among Barbarians.
C. A Barbarian who is more personally clean than another also tends to have made more raids.
D. There is an 81% chance that the relationship between personal cleanliness and number of raids is linear.
E. The variation in the residuals for the number of raids among Barbarians is about 19% of the variation in the original responses.
Investigative Tasks

AP9. Siri’s equation. Athletes and exercise scientists sometimes use the proportion of fat to overall body mass as one measure of fitness, but measuring the percentage of body fat directly poses a challenge. Fortunately, some good statistical detective work by W. E. Siri in the 1950s provided an alternative to direct measurement that is still in use today. Siri’s method lets you estimate the percentage of body fat from body density, which you can measure directly by hydrostatic (underwater) weighing. In this exercise, you’ll see how transformations and residual plots play a crucial role in finding Siri’s model.

a. A first model. Use the data in Display 3.131.

<table>
<thead>
<tr>
<th>Density</th>
<th>Percentage of Body Fat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.053</td>
<td>19.94</td>
</tr>
<tr>
<td>1.053</td>
<td>20.04</td>
</tr>
<tr>
<td>1.055</td>
<td>19.32</td>
</tr>
<tr>
<td>1.056</td>
<td>18.56</td>
</tr>
<tr>
<td>1.048</td>
<td>22.08</td>
</tr>
<tr>
<td>1.040</td>
<td>25.81</td>
</tr>
<tr>
<td>1.030</td>
<td>30.78</td>
</tr>
<tr>
<td>1.064</td>
<td>15.33</td>
</tr>
</tbody>
</table>

Display 3.131 The percentage of body fat and body density of 15 women. [Source: M. L. Pollock, University of Florida, 1956.]

i. Plot percentage of body fat versus body density, and from your plot explain whether you think a line gives a poor fit, a moderately good fit, or an extremely good fit.

ii. Write the equation of the least squares line.

iii. Does the value of $r^2$ tend to confirm your opinion about how well the line fits?

iv. Construct a residual plot and describe the pattern. Does the plot tend to confirm or raise questions about your opinion? Explain.

b. A new model.

i. Explain how knowing that fat is less dense than the rest of the body might have led Siri to plot the percentage of body fat against the reciprocal of density: $\frac{1}{\text{density}}$. Construct this plot and fit a least squares line, and compare its equation to the one Siri found:

\[
\% \text{ body fat} = -450 + 495 \left(\frac{1}{\text{density}}\right)
\]

Next, plot residuals versus $\frac{1}{\text{density}}$.

What features of this plot confirm that the transformation has improved the linear fit?

ii. Find the correlation between the percentage of body fat and body density and the correlation between the percentage of body fat and the reciprocal of body density. Comment on using correlation as the only criterion for assessing the usefulness of a model.

iii. Suggest another model for the percentage of body fat and body density data that might work nearly as well as Siri’s.
AP10. In AP9, you explored the relationship between the percentage of body mass that is fat and body density. Display 3.132 is an extension of the data in AP9, including skinfold measurements and data for men.

a. Does Siri’s model for relating percentage of body fat to body density hold for men as well as it did for women? That model was

\[ \text{% body fat} = -450 + 495 \left( \frac{1}{\text{density}} \right) \]

b. The variable skinfold is the sum of a number of skinfold thicknesses taken at various places on the body. (The units are millimeters.) The skinfold measurements are used to predict body density. Find a good model for predicting density from skinfold measurements based on these data for women. Do your models require re-expression?

<table>
<thead>
<tr>
<th>Women</th>
<th>Men</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Fat</td>
<td>Density</td>
</tr>
<tr>
<td>19.94</td>
<td>1.053</td>
</tr>
<tr>
<td>20.04</td>
<td>1.053</td>
</tr>
<tr>
<td>19.32</td>
<td>1.055</td>
</tr>
<tr>
<td>18.56</td>
<td>1.056</td>
</tr>
<tr>
<td>22.08</td>
<td>1.048</td>
</tr>
<tr>
<td>25.81</td>
<td>1.040</td>
</tr>
<tr>
<td>30.78</td>
<td>1.030</td>
</tr>
<tr>
<td>15.33</td>
<td>1.064</td>
</tr>
<tr>
<td>44.33</td>
<td>1.001</td>
</tr>
<tr>
<td>20.74</td>
<td>1.052</td>
</tr>
<tr>
<td>38.32</td>
<td>1.014</td>
</tr>
<tr>
<td>9.91</td>
<td>1.076</td>
</tr>
<tr>
<td>15.81</td>
<td>1.063</td>
</tr>
<tr>
<td>34.02</td>
<td>1.023</td>
</tr>
</tbody>
</table>

Display 3.132 Percentage of body fat, body density, and skinfold for 15 women and 14 men. [Source: M. L. Pollock, University of Florida, 1956.]
What prompts a hamster to prepare for hibernation? A student designed an experiment to see whether the number of hours of light in a day affects the concentration of a key brain enzyme.
Most of what you’ve done in Chapters 2 and 3, as well as in the first part of Chapter 1, is part of data exploration—ways to uncover, display, and describe patterns in data. Methods of exploration can help you look for patterns in just about any set of data, but they can’t take you beyond the data in hand. With exploration, what you see is all you get. Often, that’s not enough.

**Pollster:** I asked a hundred likely voters who they planned to vote for, and fifty-two of them said they’d vote for you.

**Politician:** Does that mean I’ll win the election?

**Pollster:** Sorry, I can’t tell you. My stat course hasn’t gotten to inference yet.

**Politician:** What’s inference?

**Pollster:** Drawing conclusions based on your data. I can tell you about the hundred people I actually talked to, but I don’t yet know how to use that information to tell you about all the likely voters.

Methods of inference can take you beyond the data you actually have, but only if your numbers come from the right kind of process. If you want to use 100 likely voters to tell you about all likely voters, how you choose those 100 voters is crucial. The quality of your inference depends on the quality of your data; in other words, bad data lead to bad conclusions. This chapter tells you how to gather data through surveys and experiments in ways that make sound conclusions possible.

Here’s a simple example.\(^1\) When you taste a spoonful of chicken soup and decide it doesn’t taste salty enough, that’s exploratory analysis: You’ve found a pattern in your one spoonful of soup. If you generalize and conclude that the whole pot of soup needs salt, that’s an inference. To know whether your inference is valid, you have to know how your one spoonful—the data—was taken from the pot. If a lot of salt is sitting on the bottom, soup from the surface won’t be representative, and you’ll end up with an incorrect inference. If you stir the soup thoroughly before you taste, your spoonful of data will more likely represent the whole pot. Sound methods for producing data are the statistician’s way of making sure the soup gets stirred so that a single spoonful—the sample—can tell you about the whole pot. Instead of using a spoon, the statistician relies on a chance device to do the stirring and on probability theory to make the inference.

Soup tasting illustrates one kind of question you can answer using statistical methods: Can I generalize from a small sample (the spoonful) to a larger population (the whole pot of soup)? To use a sample for inference about a population, you must *randomize*, that is, use chance to determine who or what gets into your sample.

\(^1\)The inspiration for this metaphor came from Gudmund Iversen, who teaches statistics at Swarthmore College.
The other kind of question is about comparison and cause. For example, if people eat chicken soup when they get a cold, will this cause the cold to go away more quickly? When designing an experiment to determine if a pattern in the data is due to cause and effect, you also must randomize. That is, you must use chance to determine which subjects get which treatments. To answer the question about chicken soup, you would use chance to decide which of your subjects eat chicken soup and which don't, and then compare the duration of their colds.

The first part of this chapter is about designing surveys. A well-designed survey enables you to make inferences about a population by looking at a sample from that population. The second part of the chapter introduces experiments. An experiment enables you to determine cause by comparing the effects of two or more treatments.

In this chapter, you will learn

- reasons for using samples when conducting a survey
- how to design a survey by randomly selecting participants
- how surveys can go wrong (bias)
- how to design a sound experiment by randomly assigning treatments to subjects
- how experiments can determine cause
- how experiments can go wrong (confounding)
4.1 Why Take Samples, and How Not To

Taking a sample survey can help you determine the percentage of people in a population who have a particular characteristic. For example, the Gallup poll periodically asks adults in the United States questions such as whether they approve of the job the president is doing. Polls or surveys such as this rely on samples to get their percentages; that is, they don’t ask every adult in the United States but instead ask only a sample of about 1500. Similarly, quality assurance methods in a manufacturing plant do not call for checking the quality of every item coming off the production line; rather, they recommend that a limited number of items (a sample) be checked carefully for quality.

Cost in money or time is a primary reason to use samples. Imagine: It’s Sunday night at 8:00, and Nielsen Media Research is gathering data about what proportion of TV sets are tuned to a particular program and what kinds of people are watching that program. To find out how many TV sets are tuned to the program, electronic meters have been hooked up to televisions in a sample of households. To find out who is watching it, a sample of people are filling out diaries. Why doesn’t Nielsen Media Research include everyone in the United States in these surveys? To hook up a meter to every TV set would cost more than anyone would be willing to pay for the information. Also, to try to get a diary from every TV viewer about what they were watching at 8:00 p.m. on Sunday would take so much time that the information would no longer be very useful. So for two reasons—money and time—Nielsen ratings are based on samples.

Sampling lets a cook know how the soup will taste without eating it all just to make sure. A light bulb manufacturer can’t test the life of every bulb produced, or there would be none left to sell. Whenever testing destroys the things you test, your only choice is to work with a sample.

If time and money are limited—and they always are—there’s a tradeoff between the number of people in your sample and the amount and quality of information you can expect to get from each person. Using a sample allows you to spend more time and money gathering high-quality information from each individual. This often produces greater accuracy in the results than you could get from a quick, but error-prone, study of every individual.

Census Versus Sample

In statistics, the set of people or things that you want to know about is called the population. The individual elements of the population sometimes are called units. In everyday language, population often refers to the number of units in the set, as when you say “In 1990, the population of Massachusetts was about 6 million.” In statistics, population refers to the set itself (for example, the people of Massachusetts). The number of units is called the population size. Ordinarily, you don’t get to record data on all the units in the population, so you use a sample. The sample is the set of units you do get to study. The special case where you collect data on the entire population is a census.

Nielsen Media Research takes a survey so they can get an estimate of the proportion of all U.S. households that are tuned to a particular television program.
The true proportion that Nielsen would get from a survey of every household is called a population parameter. Nielsen uses the proportion in the sample as an estimate of this parameter. Such an estimate from a sample is called a statistic.

Not all statistics are created equal. Some aren't very good estimators of the population parameter. For example, the maximum in the sample isn't a very good estimator of the maximum in the population—it is almost always too small. You will learn more about the properties of estimators in Chapter 7.

**Census Versus Sample**

D1. In which of these situations do you think a census is used to collect data, and in which do you think sampling is used? Explain your reasoning.
   a. An automobile manufacturer inspects its new models.
   b. A cookie producer checks the number of chocolate chips per cookie.
   c. The U.S. president is determined by an election.
   d. Weekly movie attendance figures are released each Sunday.
   e. A Los Angeles study does in-depth interviews with teachers in order to find connections between nutrition and health.

**Bias: A Potential Problem with Survey Data**

Samples offer many advantages, but some samples are more trustworthy than others. In this section, you will learn about two ways to get untrustworthy results:

- bias in the way you select your sample
- bias in the way you get a response from the units in your sample

In Activity 4.1a, you'll examine the kind of problem that can lie in wait for the unwary sampler. Suppose you have just won a contract to estimate the average length of stay in the children's ward of a hospital. How will you gather the data?

### ACTIVITY 4.1a Time in the Hospital

**What you'll need:** one deck of cards for you and your partner

In this activity, you'll estimate the average length of stay in a five-bed hospital ward. You'll sample from the population “lengths of stay,” represented by the numbers on the 40 cards (not counting face cards) in an ordinary deck of cards. You'll estimate the average length of stay from a sample of five patients.

1. Shuffle your 40 cards several times. Deal out a row of five cards to represent your first patients. The numbers on the cards tell you how many days they will be on your ward. For example, suppose your first five cards are

   A♥ 2♥ 8♥ 8♦ 3♣

   These cards mean the patient in bed 1 will be in the hospital for 1 day, the patient in bed 2 will be there for 2 days, the patient in bed 3 will be there for 8 days, and so forth. Your partner should record this information on a
chart like the one in Display 4.1. Place the cards that represent patients in a stack separate from unused cards.

**Display 4.1** Lengths of stay for the first five patients.

2. Deal out the other cards one at a time, assigning the next patient to a bed as it becomes available. To continue with the example, suppose the next patient is 9♣. The first available bed is bed 1, and the patient will be in it for 9 days. The next available bed is bed 2, and the next patient, say, 9♦, will be there for 9 days. Display 4.2 shows a chart of the lengths of stays at this stage. The next patient will go into bed 5.

**Display 4.2** Lengths of stay for the first seven patients.

3. Continue dealing out patient cards, each representing a hospital stay. Record these stays until all five beds have been filled for at least 20 days. You should end up with something like Display 4.3. (Save your chart; you’ll need it later.)

**Display 4.3** Lengths of stay for the patients during the first 20 days.

(continued)
4. Select a day at random from days 1 through 20 (or however many days all of your beds are full).

5. Compute the average length of stay for the five patients in the beds on that day.

6. Pool your results with those of the rest of your class until you have about 30 estimates of the average length of a stay. Make a dot plot of the 30 averages from your class. Where is this distribution centered? Is there much variability in your estimates?

7. Compute the average stay for the whole population (the original deck of all 40 cards).

8. Compare your results in step 6 and step 7. On average, are your estimates generally too low, too high, or about right? Why is this the case?

9. What are the units in this situation? Did every unit have an equal chance of being in the sample? If so, explain. If not, which units had the greater chance?

10. How could you improve the sampling method?

In everyday language, we say an opinion is “biased” if it unreasonably favors one point of view over others. A biased opinion is not balanced, not objective. In statistics, bias has a similar meaning in that a biased sampling method is unbalanced.

A sampling method is **biased** if it produces samples such that the estimate from the sample is larger or smaller, on average, than the population parameter being estimated.

There’s an important distinction here between the sample itself and the method used for choosing the sample.

**Investigator:** What makes a good sample?

**Statistician:** A good sample is representative. That is, it looks like a small version of the population. Proportions you compute from the sample are close to the corresponding proportions you would get if you used the whole population. The same is true for other numerical summaries, such as averages and standard deviations or medians and IQRs.

**Investigator:** How can I tell if my sample is representative?

**Statistician:** There’s the rub. In practice, you can’t. You can tell only by comparing your sample with the population, and if you know that much about the population, why bother to take a sample?
Investigator: Great! First you tell me my sample should be representative, and then you tell me there's no way to know whether it is. Is that the best statisticians can do?

Statistician: Nope. Although you can't tell about any particular sample, it is possible to tell whether a sampling method is good or not. That's where bias comes in.

Investigator: I thought “biased” was just a fancy word for “nonrepresentative.” Not true?

Statistician: Now we’re getting to the point. Bias refers to the method, not the samples you get from it. A sampling method is biased if it tends to give nonrepresentative samples.

Investigator: Now I get it. I may not be able to tell whether my sample is representative, but if I use an unbiased method, then I can be confident that my sample is likely to be representative. Right?

Statistician: Now you’re thinking like a statistician. There’s more detail to come, but you have the big picture in focus.

DISCUSSION

Bias

D2. Explain the difference between “nonrepresentative” and “biased” as these terms pertain to sampling.

D3. Which statements describe an event that is possible? Which describe an event that is impossible?
   A. A representative sample results from a biased sample-selection method.
   B. A nonrepresentative sample results from a biased sample-selection method.
   C. A representative sample results from an unbiased sample-selection method.
   D. A nonrepresentative sample results from an unbiased sample-selection method.

Sample Selection Bias

Sample selection bias, or sampling bias, is present in a sampling method if samples tend to result in estimates of population parameters that systematically are too high or too low. Various forms of this selection bias can undermine the usefulness of samples and surveys.

You explored one kind of sampling bias in Activity 4.1a, in which patients who spent more days in the hospital were more likely to be selected for the sample. In fact, the chance of selection is proportional to the length of stay. A 5-day stay is five times as likely to be chosen as a 1-day stay. This type of sample selection bias is called size bias. Suppose a wildlife biologist samples lakes in a state by dropping grains of rice at random onto a map of the state and then selects for study the lakes that have rice on them. This is another example of size bias.
When a television or radio program asks people to call in and take sides on some issue, those who care about the issue will be overrepresented, and those who don't care as much might not be represented at all. The resulting bias from such a volunteer sample is called voluntary response bias and is a second type of sample selection bias.

Here's a simple sampling method: Take whatever's handy. For example, what percentage of the students in your graduating class plan to go to work immediately after graduation? Rather than find a representative sample of your graduating class, it would be a lot quicker to ask your friends and use them as your sample—quicker and more convenient, but almost surely biased because your friends are likely to have somewhat similar plans. A convenience sample is one in which the units chosen are those that are easy to include. The likelihood of bias makes convenience samples about as worthless as voluntary response samples.

Because voluntary response sampling and convenience sampling tend to be biased methods, you might be inclined to rely on the judgment of an expert to choose a sample that he or she considers representative. Such samples, not surprisingly, are called judgment samples. Unfortunately, though, experts might overlook important features of a population. In addition, trying to balance several features at once can be almost impossibly complicated. In the early days of election polling, local "experts" were hired to sample voters in their locale by filling certain quotas (so many men, so many women, so many voters over the age of 40, so many employed workers, and so on). The poll takers used their own judgment as to whom they selected for the poll. It took a very close election (the 1948 presidential election, in which polls were wrong in their prediction) for the polling organizations to realize that quota sampling was a biased method.

An unbiased sampling method requires that all units in the population have a known chance of being chosen, so you must prepare a “list” of population units, called a sampling frame or, more simply, frame, from which you select the sample. If you think about enough real examples, you’ll come to see that making this list is not something you can take for granted. For the Westvaco employees in Chapter 1 or for the 50 U.S. states, creating the list is not hard, but other populations can pose problems. How would you list all the people using the Internet worldwide or all the ants in Central Park or all the potato chips produced in the United States over a year? For all practical purposes, you can’t. There will often be a difference between the population—the set of units you want to know
about—and the sampling frame—the list of units you use to create your sample. A sample might represent the units in the frame quite well, but how well your sample represents the population depends on how well you’ve chosen your frame. Quite often, a convenient frame fails to cover the population of interest (using a telephone directory to sample residents of a neighborhood, for example), and a bias is introduced by this incomplete coverage. If you start from a bad frame, even the best sampling methods can’t save you: Bad frame, bad sample.

**Sample Selection Bias**

D4. Identify the type of sampling method used in each of these surveys. Would you expect the estimate of the parameter to be too high or too low?

a. You use your statistics class to estimate the percentage of students in your school who study at least 2 hours a night.

b. You send a survey to all people who have graduated from your school in the past 10 years. You use the mean annual income of those who reply to estimate the mean annual income of all graduates of your school in the past 10 years.

c. A study was designed to estimate how long people live after being diagnosed with dementia. The researchers took a random sample of the people with dementia who were alive on a given day. The date the person had been diagnosed was recorded, and after the person died the date of death was recorded.

D5. You want to know the percentage of voters who favor state funding for bilingual education. Your population of interest is the set of people likely to vote in the next election. You use as your frame the phone book listing of residential telephone numbers. How well do you think the frame represents the population? Are there important groups of individuals who belong to the population but not to the frame? To the frame but not to the population? If you think bias is likely, identify what kind of bias and how it might arise.

**Response Bias**

In all the examples so far, bias has come from the method of taking the sample. Unfortunately, bias from other sources can contaminate data even from well-chosen sampling units.

Perhaps the worst case of faulty data is no data at all. It isn’t uncommon for 40% of the people contacted to refuse to answer a survey. These people might be different from those who agree to participate. An example of this nonresponse bias came from a controversial study that found that left-handers died, on average, about 9 years earlier than right-handers. The investigators sent questionnaires to the families of everyone listed on the death certificates in two counties near Los Angeles asking about the handedness of the person who had died. One critic noted that only half the questionnaires were returned. Did that change the results? Perhaps. [Source: “Left-Handers Die Younger, Study Finds,” Los Angeles Times, April 4, 1991.]

Nonresponse bias, like bias that comes from the sampling method, arises from who replies. **Questionnaire bias** arises from how you ask the questions.
The opinions people give can depend on the tone of voice of the interviewer, the appearance of the interviewer, the order in which the questions are asked, and many other factors. But the most important source of questionnaire bias is the wording of the questions. This is so important that those who report the results of surveys should always provide the exact wording of the questions.

For example, Reader’s Digest commissioned a poll to determine how the wording of questions affected people’s opinions. The same 1031 people were asked to respond to these two statements:

1. I would be disappointed if Congress cut its funding for public television.
2. Cuts in funding for public television are justified as part of an overall effort to reduce federal spending.

Note that agreeing with the first statement is pretty much the same as disagreeing with the second. However, 54% agreed with the first statement, 40% disagreed, and 6% didn’t know, while 52% agreed with the second statement, 37% disagreed, and 10% didn’t know. [Source: Fred Barnes, “Can You Trust Those Polls?” Reader’s Digest, July 1995, pp. 49–54.]

Another problem polls and surveys have is trying to ensure that people tell the truth. Often, the people being interviewed want to be agreeable and tend to respond in the way they think the interviewer wants them to respond. Newspaper columnist Dave Barry reported that he was called by Arbitron, an organization that compiles television ratings. Dave reports:

So I figured the least I could do, for television, was be an Arbitron household. This involves two major responsibilities:

1. Keeping track of what you watch on TV.
2. Lying about it.

At least that’s what I did. I imagine most people do. Because let’s face it: Just because you watch a certain show on television doesn’t mean you want to admit it. [Source: Dave Barry, Dave Barry Talks Back, copyright © 1991 by Dave Barry. Used by permission of Crown Publishers, a division of Random House, Inc.]
Bias from incorrect responses might be the result of intentional lying, but it is more likely to come from inaccurate measuring devices, including inaccurate memories of people being interviewed in self-reported data. Patients in medical studies are prone to overstate how well they have followed the physician’s orders, just as many people are prone to understate the amount of time they actually spend watching TV. Measuring the heights of students with a meterstick that has one end worn off leads to a measurement bias, as does weighing people on a bathroom scale that is adjusted to read on the light side.

**Response Bias**

D6. Like Dave Barry, people generally want to appear knowledgeable and agreeable, and they want to present a favorable face to the world. How might that affect the results of a survey conducted by a school on the satisfaction of its graduates with their education?

D7. Another part of the Reader’s Digest poll described on page 226 asked Americans if they agree with the statement that it is not the government’s job to financially support television programming. The poll also asked them if they’d be disappointed if Congress cut its funding for public television. Which question do you think brought out more support for public television?

D8. How is nonresponse bias different from voluntary response bias?

**Summary 4.1: Why Take Samples, and How Not To**

The population is the set of units you want to know about. The sample is the set of units you choose to examine. A census is an examination of all units in the entire population. Important reasons for using a sample in many situations rather than taking a census include these:

- Testing sometimes destroys the items.
- Sampling can save money.
- Sampling can save time.
- Sampling can make it possible to collect more or better information on each unit.
A sampling method is biased if it tends to give results that, on average, are too low or too high. This can happen if the method of taking the sample or the method of getting a response is flawed.

Sources of bias from the method of taking the sample include

- using a method that gives larger units a bigger chance of being in the sample (size bias)
- letting people volunteer to be in the sample
- using a sample just because it is convenient
- selecting the sampling units based on “expert” judgment
- constructing an inadequate sampling frame

Types of bias derived from the method of getting the response from the sample include

- nonresponse bias
- questionnaire bias
- incorrect response or measurement bias

Practice

Census Versus Sample

P1. You want to estimate the average number of TV sets per household in your community.
   a. What is the population? What are the units?
   b. Explain the advantages of sampling over conducting a census.
   c. What problems do you see in carrying out this sample survey?

Bias

P2. Four people practicing shooting a bow and arrow made these patterns on their targets.

![Display 4.4 Results of four archers.](image)

a. Which person had shots that were biased and had low variability?

b. Which person had shots that were biased and had high variability?

c. Which person had shots that were unbiased and had low variability?

d. Which person had shots that were unbiased and had high variability?

e. Do you think it would be easiest to help Al, Cal, or Dal improve?

Sample Selection Bias

P3. Describe the type of sample selection bias that would result from each of these sampling methods.

a. A county official wants to estimate the average size of farms in a county in Iowa. He repeatedly selects a latitude and longitude in the county at random and places the farms at those coordinates in his sample. If something other than a farm is at the coordinates, he generates another set of coordinates.

b. In a study about whether valedictorians “succeed big in life,” a professor “traveled across Illinois, attending high school graduations and selecting 81 students to participate. . . . He picked students from the most diverse communities possible, from little rural schools to rich suburban schools near Chicago to city schools.”

c. To estimate the percentage of students who passed the most recent AP Statistics Exam, a teacher on an Internet discussion list for teachers of AP Statistics asks teachers on the list to report to him how many of their students took the test and how many passed.

d. To estimate the average length of the pieces of string in a bag, a student reaches in, mixes up the strings, selects one, mixes them up again, selects another, and so on.

e. In 1984, Ann Landers conducted a poll on the marital happiness of women by asking women to write to her.

P4. Suppose the Museum of Fine Arts in Boston wants to estimate what proportion of people who come to Boston from out of town planned their trip to Boston mainly to visit the museum. The sample will consist of all out-of-town visitors to the museum on several randomly selected days. On buying a ticket to the museum, people will be asked whether they came from out of town and, if so, what the main reason for their trip was. Do you expect the museum's estimate to be too high, too low, or just about right? Why? What kind of sampling method is this?

Response Bias

P5. Consider this pair of questions related to gun control:

I. Should people who want to buy guns have to pass a background check to make sure they have not been convicted of a violent crime?

II. Should the government interfere with an individual's constitutional right to buy a gun for self-defense?

Which question is more likely to show a higher percentage of people who favor some control on gun ownership?

P6. In one study, educators were asked to rank Princeton's undergraduate business program. Every educator rated it in the top ten in the country. Princeton does not have an undergraduate business program. What kind of bias is shown in this case? [Source: Anne Roark, “Guidebooks to Colleges Get A's, F's,” Los Angeles Times, November 21, 1982, Part I, p. 25.]

Exercises

E1. Suppose you want to estimate the percentage of U.S. households have children under the age of 5 living at home. Each weekday from 9 a.m. to 5 p.m., your poll takers call households in your sample. Every time they reach a person in one of the homes, they ask, “Do you have children under the age of 5 living in your household?” Eventually you give up on the households that cannot be reached.

a. Will your estimate of the percentage of U.S. households with children under the age of 5 probably be too low, too high, or about right?

b. How does this example help explain why poll takers are likely to call at dinnertime?

c. Is this a case of sampling bias or of response bias?

E2. A wholesale food distributor has commissioned a sample survey to estimate the satisfaction level of his customers, who are the owners of small restaurants. The sampling firm takes a random sample from the current list of customers, develops a satisfaction questionnaire, and sends field workers out to interview the owners of the sampled restaurants. The field workers realize that the time does not allow them to interview all the owners selected for the sample, so if the owner is busy when the field worker arrives, the worker moves on to the next business.

a. What is the population? The sample?

b. What kind of bias does this survey have?

c. What kind of restaurants would you expect to be underrepresented in the sample?
E3. Suppose you want to estimate the average response to the question “Do you like math?” on a scale from 1 (“Not at all”) to 7 (“Definitely”) for all students in your school. You use your statistics class as a sample. What kind of sample is this? What sort of bias, if any, would be likely? Be as specific as you can. In particular, explain whether you expect the sample average to be higher or lower than the population average.

E4. At a meeting of local Republicans, the organizers want to estimate how well their party’s candidate will do in their district in the next race for Congress. They use the people present at their meeting as their sample. What kind of sample is this? What bias do you expect?

E5. You want to estimate the average number of states that people living in the United States have visited. If you asked only people at least 40 years old, would you expect the estimate to be too high or too low? What bias might you expect if you take your sample only from those living in Rhode Island?

E6. For a study on smoking habits, you want to estimate the proportion of adult males in the United States who are nonsmokers, who are cigarette smokers, and who are pipe or cigar smokers. Tell why it makes more sense to use a sample than to try to survey every individual. What types of bias might show up when you attempt to collect this information?

E7. “Television today is more offensive than ever, say the overwhelming majority—92%—of readers who took part in USA Weekend’s third survey measuring attitudes toward the small screen.” More than 21,600 people responded to this write-in survey. [Source: USA Weekend, May 16–18, 1997, p. 20.]

a. What kind of sample is this?

b. Do you trust the results of the survey? Why or why not?

c. What percentage of the entire U.S. television-watching public do you think would say that today’s shows are more offensive than ever: more than 92%, quite a bit less than 92%, or just about 92%? Why?

E8. Are people willing to change their driving habits in the face of higher gasoline prices? At a time of steeply rising gas prices, a Consumers Union poll of Internet users who chose to participate showed these results. [Source: www.consumersunion.org.]

<table>
<thead>
<tr>
<th></th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Made no changes</td>
<td>42%</td>
</tr>
<tr>
<td>Driven less</td>
<td>53%</td>
</tr>
<tr>
<td>Carpoled</td>
<td>2%</td>
</tr>
<tr>
<td>Relied more on mass transportation</td>
<td>3%</td>
</tr>
</tbody>
</table>

**Total Votes: 139**

a. What kind of sample is this?

b. Do you trust the results of this survey? Why or why not?

c. What percentage of U.S. drivers do you think drive less as a result of higher gasoline prices: more than 53%, less than 53%, or just about 53%? Why do you think that?

E9. To estimate the average number of children per family in the city where you live, you use your statistics class as a convenience sample. You ask each student in the sample how many children are in his or her family. Do you expect the sample average to be higher or lower than the population average? Explain why.

E10. Suppose you wish to estimate the average size of English classes on your campus. Compare the merits of these two sampling methods.

I. You get a list of all students enrolled in English classes, take a random sample of those students, and find out how many students are enrolled in each sampled student’s English class.

II. You get a list of all English classes, take a random sample of those classes, and find out how many students are enrolled in each sampled class.
E11. A Gallup/CNN/USA Today poll asked this question:
As you may know, the Bosnian Serbs rejected the United Nations peace plan and Serbian forces are continuing to attack Muslim towns. Some people are suggesting that the United States conduct air strikes against Serbian military forces, while others say we should not get militarily involved. Do you favor or oppose U.S. air strikes?
On the same day, the ABC News poll worded its question this way:
Specifically, would you support or oppose the United States, along with its allies in Europe, carrying out air strikes against Bosnian Serb artillery positions and supply lines?
Explain which poll you think got a larger response favoring air strikes.

E12. Using the gun control example in P5 on page 229 as a guide, write two versions of a question about a controversial issue, with one version designed to get a higher percentage in favor and the other designed to get a higher percentage opposed. Test your questions on a sample of 20 people.

E13. Display 1.1 (page 5) lists data on the 50 employees of the engineering department at Westvaco. How would you take a sample of 20 employees that is representative of all 50 with regard to all the features listed: pay type, year of birth, year of hire, and whether laid off?

E14. Bring to class an example of a survey from the media. Identify the population and the sample, and discuss why a sample was used rather than a census. Do you see any possible bias in the survey method?

4.2 Random Sampling: Playing It Safe by Taking Chances

The best-planned surveys leave a lot to chance. The key idea is to randomize, that is, let chance choose the sampling units. This might seem like a paradox at first, but it makes sense once you understand a basic fact about chancelike behavior: Selecting your sample by chance is the only method guaranteed to be unbiased.

You’ve seen how some common sampling methods are biased, tending to give nonrepresentative samples. People who care enough to respond voluntarily might not be typical of others. The members of your class, although they’re convenient to survey, don’t represent some groups well at all. Even relying on expert judgment doesn’t work as well as we’d like. Over the long run, chance beats all these methods.

Simple Random Samples

When each unit in the population has a fixed probability of ending up in the sample, we call the sample a probability sample. One type of probability sample is a simple random sample, or SRS.

In simple random sampling, all possible samples of a given fixed size are equally likely. All units have the same chance of belonging to the sample, all possible pairs of units have the same chance of belonging to the sample, all possible triples of units have the same chance, and so on.
The simplest way to let chance choose your simple random sample amounts to putting all the individuals from your population into a gigantic hat, mixing them thoroughly, and drawing individuals out one at a time until you have enough for your sample. (This is like mixing the soup before tasting.) Although this method is exactly right in theory, in practice mixing thoroughly is almost impossible. So it's actually better to implement the theory using random numbers.

**Steps in Choosing a Simple Random Sample**

1. Start with a list of all the units in the population.
2. Number the units in the list.
3. Use a random number table or generator to choose units from the numbered list, one at a time, until you have as many as you need.

Do you think you can select representative sampling units as well as a random number table can? Activity 4.2a provides an opportunity for you to test yourself.

**ACTIVITY 4.2a Random Rectangles**

**What you’ll need:** Display 4.5, a method of producing random digits

Your goal is to choose a sample of 5 rectangles from which to estimate the average area of the 100 rectangles in Display 4.5.

1. Without studying the display of rectangles too carefully, quickly choose five that you think represent the population of rectangles on the page. This is your judgment sample.
2. Find the areas of each rectangle in your sample of five and compute the sample mean, that is, the average area of the rectangles in your sample.
3. List your sample mean with those of the other students in the class. Construct a plot of the means for your class.
4. Describe the shape, center, and spread of the plot of sample means from the judgment samples.

(continued)
Display 4.5  Random rectangles.
5. Now generate five distinct random integers between 0 and 99. (The rectangle numbered 100 can be called 0.) Find the rectangles whose numbers correspond to your random numbers. This is your random sample of five rectangles. [You can use a calculator to generate random numbers. See Calculator Note 4A.]

6. Repeat steps 2–4, this time using your random sample.
7. Discuss how the two distributions of sample means are similar and how they differ.
8. Which method of producing sample means do you think is better if the goal is to use the sample mean to estimate the population mean? (Your instructor has the value of the population mean.)

**Simple Random Samples**

D9. Use a table of random digits to take a simple random sample of six students from your class. For what characteristics is the sample representative of this class? Not representative?

D10. Describe how you could obtain a simple random sample of all students enrolled in the English classes at your school.

D11. **Readability.** You’ve decided you simply must know the proportion of capital letters in this textbook because it can indicate the complexity of the sentences on a page. Think of the book itself as your population and each printed character as a unit. How could you get a simple random sample of 10,000 characters? What are the advantages and disadvantages of this sampling method?

**Stratified Random Samples**

Suppose you are planning a sample survey for an international outdoor clothing manufacturer to see if their image has suffered among their retailers due to negative press coverage regarding their use of sweatshop labor. The company headquarters furnishes you with a list of retail outlets throughout the world that sell their products. Because press coverage is largely national, it will be more informative to organize this list by country and take a random sample from each country. You’ll be gathering data using each country’s phone or postal system in the language of that country, so categorizing the retail outlets by country will also be more convenient.
Classifying the retail outlets by country is called **stratification**. If you can divide your population into subgroups that do not overlap and that cover the entire sampling frame, the subgroups are called **strata**. If the sample you take within each subgroup is a simple random sample, your result is a **stratified random sample**. This is a second type of probability sample.

### Steps in Choosing a Stratified Random Sample

1. Divide the units of the sampling frame into non-overlapping subgroups.
2. Take a simple random sample from each subgroup.

Why stratify?

Why might you want to stratify a population? Here are the three main reasons:

- **Convenience** in selecting the sampling units is enhanced. It is easier to sample in smaller, more compact groups (countries) than in one large group spread out over a huge area (the world).
- **Coverage** of each stratum is assured. The company might want to have data from each country in which it sells products; a simple random sample from the frame does not guarantee that this would happen.
- **Precision** of the results may be improved. That is, stratification tends to give estimates that are closer to the value for the entire population than does an SRS. This is the fundamental statistical reason for stratification.

An example might help clarify the last point.

**Example: To Stratify or Not to Stratify**

Suppose a geologist has a box of rocks from a mountain stream in Mexico. Among other things, she wants to estimate the mean diameter of these discus-shaped rocks so it can be compared to the mean diameter of similar rocks from other locations. She has a sieve with a 2-inch mesh, which allows through only rocks under 2 inches in diameter. To sieve or not to sieve before estimating, that is the question. In other words, is it better to stratify the rocks into two groups (small and large) and then sample from each stratum, or is it just as good to sample from the pile of rocks without sieving?

To simplify the problem, consider a population of 100 rocks. After sieving, 50 turn out to have diameters of less than 2 inches. Display 4.6 shows dot plots of the population of diameters and of that population divided into the two strata (small and large rocks).

Twenty simple random samples, each of size 4, were taken from the population, and the sample mean was calculated for each. A plot of the 20 sample means is shown in the top half of Display 4.7. Next, 20 stratified random samples were taken, with each stratified sample consisting of a simple random sample of size 2 from the small rocks mixed with a simple random sample of size 2 from the large rocks. A plot of these 20 sample means is shown in the bottom half of Display 4.7. Note that for both sampling methods we have 20 samples, each of size 4. Which method is better?
Display 4.6 Data on rock diameters, population and strata.

Display 4.7 Means of simple and stratified random samples of rock diameters.

Solution

Each method is unbiased because the sample means are centered at the population mean, about 2.51. However, the means from the stratified random samples are less variable, tending to lie closer to the population mean. Because it’s better to have estimates that are closer to the parameter than estimates that tend to be farther away, stratification wins.

From this example, it certainly looks as though the stratification (sieving) pays off in producing estimates of the mean with smaller variation. This will be true generally if the stratum means are quite different. If the geologist had decided to stratify based on color and if color was not related to diameter, then the stratification would have produced little or no gain in the precision of the estimates. The guiding principle is to choose strata that have very different means, whenever that is possible.

One good way to choose the relative sample sizes in stratified random sampling is to make them proportional to the stratum sizes (the number of units in the stratum). Thus, if all the strata are of the same size, the samples should all

Make the strata as different as possible.

Allocate units in the sample proportionally to the number of units in the strata.
be of the same size. If a population is known to have 65% women and 35% men, then a sample of 100 people stratified on gender should contain 65 women and 35 men. If the sample sizes are proportional to the stratum sizes, then the overall sample mean (the mean calculated from the samples of all the strata mixed together) will be a good estimate of the population mean. Proportional allocation is particularly effective in reducing the variation in the sample means if the stratum standard deviations are about equal. (If the stratum standard deviations differ greatly, a more effective allocation of samples to strata can be found.)

DISCUSSION

Stratified Random Samples

D12. Suppose the geologist had a set of sieves that could divide the rocks into many small piles with essentially equal-diameter rocks in each pile. Would this process result in a more precise estimate of the mean diameter than that of the two-strata scenario discussed earlier? Why or why not?

D13. An administrator wants to estimate the average amount of time high school students spend traveling to school. The plan is to stratify the students according to grade level and then take a simple random sample from each grade. What is potentially good and what is potentially bad about this plan?

D14. Your assignment is to estimate the mean number of hours per week spent studying by students in your school. Discuss how you would set up a stratified random sampling plan to accomplish the task.

Other Methods of Sampling

Simple random samples and stratified random samples offer many advantages, but you often run into a major practical problem in attempting to carry them out. Sampling individual units from a population one at a time is often too costly, too time consuming, or, if a good frame is not available, simply not possible. Several commonly used sampling designs get around this problem by forming larger sampling units out of groups of population units. (Think of sampling households in a neighborhood rather than individual residents.) If the groups are chosen wisely, these designs are often nearly as good as simple or stratified random sampling.

Cluster Samples

To see how well U.S. 4th graders do on an arithmetic test, you might take a simple random sample of children enrolled in the 4th grade and give each child a standardized test. In theory, this is a reasonable plan, but it is not very practical. For one thing, how would you go about making a complete list of all the 4th graders in the United States? For another, imagine the work required to track down each child in your sample and get him or her to take the test. Instead of taking an SRS of 4th graders, it would be more realistic to take an SRS of elementary schools and then give the test to all the 4th graders in those schools. Getting a list of all the elementary schools in the United States is a lot easier than getting a list of all the individuals enrolled in the 4th grade. Moreover, once you’ve chosen your sample of elementary schools, it’s a relatively easy organizational problem to give the test to all 4th graders in those schools. This is an example of
cluster sampling, with each elementary school a cluster of 4th-grade students. A cluster sample is an SRS of non-overlapping clusters of units.

**Steps in Choosing a Cluster Sample**

1. Create a numbered list of all the clusters in your population.
2. Take a simple random sample of clusters.
3. Obtain data on each individual in each cluster in your SRS.

The situation of the 4th graders illustrated the two main reasons for using cluster samples: You need only a list of clusters rather than a list of individuals, and for some studies it is much more efficient to gather data on individuals grouped by clusters than on all individuals one at a time.

**Two-Stage Cluster Samples**

The National Assessment of Educational Progress (NAEP) uses a variation on cluster sampling. For example, to get a sample of 4th graders in Illinois, NAEP takes a (stratified) random sample of elementary schools. Then, typically 30 4th-grade students are selected randomly from each school to take the mathematics test. [Source: nces.ed.gov] This is an example of two-stage sampling because it involves two randomizations.

**Steps in Choosing a Two-Stage Cluster Sample**

1. Create a numbered list of all the clusters in your population, and then take a simple random sample of clusters.
2. Create a numbered list of all the individuals in each cluster already selected, and then take an SRS from each cluster.

Two-stage cluster samples are useful when it is much easier to list clusters than individuals but still reasonably easy and sufficient to sample individuals once the clusters are chosen. Two-stage cluster sampling might sound like stratified random sampling, but they are different.

- In stratified random sampling, you want to choose strata so that the units within each stratum don’t vary much from each other. Then you sample from every stratum.
- In two-stage cluster sampling, you want to choose clusters so that the variation within each cluster reflects the variation in the population, if possible. Then you sample from within only some of your clusters.

**Systematic Samples with Random Start**

To get a quick sample of the students in your class, use the common system for choosing teams by “counting off.” Count off by, say, eights, and then select a digit between 1 and 8 at random. Every student who calls out that digit is in the sample. This is an example of a systematic sample with random start.
4.2 Random Sampling: Playing It Safe by Taking Chances

Steps in Choosing a Systematic Sample with Random Start

1. By a method such as counting off, divide your population into groups of the size you want for your sample.
2. Use a chance method to choose one of the groups for your sample.

Suppose your population units are in a list, such as a list of names in a directory, and you want to select a sample of a certain size. The steps in choosing a systematic sample with random start are equivalent to choosing a random starting point between units 1 and \( k \) and then taking every \( k \)th unit thereafter. The value of \( k \) is determined by the sample size. For example, suppose you want a systematic sample of 20 units selected from a list of 1000 units. Then \( k \) is 50, because selecting a random start between 1 and 50 and then taking every 50th unit thereafter will result in a systematic sample of 20 units.

When the units in the population are well mixed before the counting off, this method will produce a sample much like an SRS. Systematic samples, like cluster samples, are often easier to take than simple random samples of the same size.

Other Methods of Sampling

D15. Why isn’t taking a systematic sample with random start equivalent to taking a simple random sample?

D16. The *Los Angeles Times* commissioned an analysis of St. John’s wort, an over-the-counter herbal medicine, to determine whether the potency of the pills matched the potency claimed on the bottle. The sampling procedure was described as this: “For the analysis, 10 pills were sampled from each of three containers of one lot of each product.” [Source: *Los Angeles Times*, August 31, 1998, p. A10.]

a. What type of sampling was this?
b. Can you suggest an improvement in the sampling procedure?

D17. A newspaper reports that about 60% of the cars in your community come from manufacturers based outside the United States. How would you design a sampling plan for gathering the data to substantiate or refute this claim?

D18. Both cluster sampling and stratified random sampling involve viewing the sampling frame as a collection of subgroups. Explain the difference between these two types of sampling.

D19. Take a systematic sample of five students from your class.

Summary 4.2: Random Sampling

The main reason for using a chance method to choose the individuals for your sample is that randomization protects against sample selection bias. And, as you will see, random selection of the sampling units will make inference about the population possible.

There are several ways to take a sample using random selection. Some of the most common are described on the next page.
• **Simple random sample.** Number the individuals and use random numbers to select those to be included in the sample.

• **Stratified random sample.** First divide your population into groups, called strata, and then take a simple random sample from each group.

• **Cluster sample.** First select clusters at random, and then use all the individuals in the clusters as your sample.

• **Two-stage cluster sample.** First select clusters at random, and then select individuals at random from each cluster.

• **Systematic sample with random start.** “Count off” your population by a number \( k \) determined by the size of your sample, and select one of the counting numbers between 1 and \( k \) at random; the units with that number will be in your sample.

Display 4.8 shows schematically how five common sampling designs might look for sampling the blobs in the rectangles.
Simple Random Samples

P7. Decide whether these sampling methods produce a simple random sample of students from a class of 30 students. If not, explain why not.

a. Select the first six students on the class roll sheet.

b. Pick a digit at random and select those students whose phone number ends in that digit.

c. If the classroom has six rows of chairs with five seats in each row, choose a row at random and select all students in that row.

d. If the class consists of 15 males and 15 females, assign the males the numbers from 1 to 15 and the females the numbers from 16 to 30. Then use a random digit table to select six numbers from 1 to 30. Select the students assigned those numbers for your sample.

e. If the class consists of 15 males and 15 females, assign the males the numbers from 1 to 15 and the females the numbers from 16 to 30. Then use a random digit table to select three numbers from 1 to 15 and three numbers from 16 to 30. Select the students assigned those numbers for your sample.

f. Randomly choose a letter of the alphabet and select for the sample those students whose last name begins with that letter. If no last name begins with that letter, randomly choose another letter of the alphabet.

Stratified Random Samples

P8. Receding gums are most common in people over age 40, but they can occur in younger adults who have periodontal (gum) disease. There does not appear to be a great difference between the rates of males and females with receding gums. You have been hired to estimate the percentage of all adults in your city with receding gums and were given a budget sufficient for a random sample of 1200 adults.

a. Suppose that, in your city, 43% of adults see their dentist twice a year, 32% once a year, and 25% less frequently than once a year. If you stratify on number of visits to a dentist, how many adults from each stratum should you include in your sample?

b. In addition to number of visits to a dentist, which of these variables would be the best to consider stratifying on? Explain your answer.

- gender (male/female)
- age (over 40/under 40)
- whether a person eats an apple a day

Other Methods of Sampling

P9. Suppose your population is 65% women and 35% men. In a stratified random sample of 100 women and 100 men, you find that 84 women pump their own gas and 69 men pump their own gas. What is the best estimate of the proportion of the entire population who pump their own gas?

P10. Suppose 200 people are waiting in line for tickets to a rock concert. You are working for the school newspaper and want to interview a sample of the people in line. Show how to select a systematic sample of

a. 5% of the people in line.

b. 20% of the people in line.

P11. The American Statistical Association directory lists its roughly 17,000 members in alphabetical order. You want a sample of about 1000 members. Describe how to use the alphabetical listing to take a systematic sample with random start.
P12. A bank keeps its list of home mortgages in chronological order, from earliest to latest in terms of the date they were assumed. An auditor wants to get a quick check on the average balance of these mortgages by taking a sample.
   a. Describe how the bank could take a systematic sample with random start.
   b. Describe how to take a cluster sample.
   c. Describe how to take a two-stage cluster sample.

   a. Instead of an SRS, consider taking a cluster sample. What would you use for your clusters? What are the advantages of the cluster sample? Are there any disadvantages?
   b. Suppose you want to estimate the average number of capital letters per printed line, so now your sampling unit is a line. Describe how to take a two-stage cluster sample from this textbook.
   c. Now suppose your sampling unit is an individual printed character. Describe how to take a three-stage cluster sample of characters from this book.

Exercises

E15. A wholesale food distributor has hired you to conduct a sample survey to estimate the satisfaction level of the businesses he serves, which are mainly small grocery stores and restaurants. A current list of businesses served by the distributor is available for the selection of sampling units.
   a. If the distributor wants good information from both the small grocery store owners and the restaurant owners, what kind of sampling plan will you use?
   b. If the distributor wants information from the customers who frequent the grocery stores and restaurants he serves, how would you design the sampling plan?

E16. Cookies. Which brand of chocolate chip cookies gives you the most chips per cookie? For the purpose of this question, take as your population all the chocolate chip cookies now in the nearest supermarket. Each cookie is a unit in this population.
   a. Explain why it would be hard to take an SRS.
   b. Describe how to take a cluster sample of chocolate chip cookies.
   c. Describe how you would take a two-stage cluster sample. What circumstances would make the two-stage cluster sample better than the cluster sample?

E17. Haircut prices. You want to take a sample of students in your school in order to estimate the average amount they spent on their last haircut. Which sampling method do you think would work best: a simple random sample; a stratified random sample with two strata, males and females; or a stratified random sample with class levels as strata? Give your reasoning.

E18. The Oxford Dictionary of Quotations, 3rd edition, has about 600 pages of quotations. Describe how you would take a systematic sample of 30 pages to use for estimating the number of typographical errors per page.

E19. An early use of sampling methods was in crop forecasting, especially in India, where an accurate forecast of the jute yield in the 1930s made some of the techniques (and their inventors) famous. Your job is to estimate the total corn yield, right before harvest, for a county with five farms and a total of 1000 acres planted in corn. How would you do the sampling?

E20. You are called upon to advise a local movie theater on designing a sampling plan for a survey of patrons on their attitudes about recent movies. About 64% of the patrons are adults, 30% are teens, and 6% are children. The theater has the time and money to
interview about 50 patrons. What design would you use?

E21. Quality-control plans in industry often involve sampling items for inspection. A manufacturer of electronic relays (switches) for the TV industry wants to set up a quality-control sampling plan for these relays as they come off its production line. What sampling plan would you suggest if there is only one production line? If there are five production lines?

E22. Suppose that a sample of 25 adults contains only women. Two explanations are possible: (1) The sampling procedure wasn’t random, or (2) a nonrepresentative sample occurred just by chance. In the absence of additional information, which explanation would you be more inclined to believe?

4.3 Experiments and Inference About Cause

The previous two sections showed you how to collect data that will allow you to generalize or infer from the sample you see to a larger but perhaps unseen population. A second kind of inference, maybe even more important than the first, takes you from a pattern you observe to a conclusion about what caused the pattern. Does regular exercise cause your heart rate to go down? Does bilingual education increase the percentage of non-native-English-speaking students who graduate from high school? Will taking a special preparatory course raise your SAT scores? Does smoking cigarettes cause lung cancer?

Cause and Effect

Children who drink more milk have bigger feet than children who drink less milk. There are three possible explanations for this association:

I. Drinking more milk causes children's feet to be bigger.

II. Having bigger feet causes children to drink more milk.

III. A lurking variable is responsible for both.

In this case, it is explanation III. The lurking variable—one that lies in the background and may or may not be apparent at the outset but, once identified, could explain the pattern between the variables—is the child's overall size. Bigger
children have bigger feet, and they drink more milk because they eat and drink more of everything than do smaller children.

But suppose you think that explanation I is the reason. How can you prove it? Can you take a bunch of children, give them milk, and then sit and wait to see if their feet grow? That won't prove anything, because children's feet will grow whether they drink milk or not.

Can you take a bunch of children, randomly divide them into a group that will drink milk and a group that won't drink milk, and then sit and wait to see if the milk-drinking group grows bigger feet? Yes! Such an experiment is just about the only way to establish cause and effect.

A Real Experiment: Kelly’s Hamsters

The goal of every experiment is to establish cause by comparing two or more conditions, called treatments, using an outcome variable, called the response. Here’s an example of a real experiment, planned and carried out by Kelly Acampora as part of her senior honors project in biology at Mount Holyoke College. What happens when an animal gets ready to hibernate? This question is too general to answer with a single experiment, but if you know a little biology, you can narrow the question enough for it to serve as the focus of an experiment. Kelly relied on three previously known facts:

2. Hamsters rely on the amount of daylight to trigger hibernation.
3. An animal’s capacity to transmit nerve impulses depends in part on an enzyme called Na\(^+\)K\(^+\) ATP-ase.

Here are the components of Kelly’s experimental design:

**Kelly’s Question:** If you reduce the amount of light a hamster gets from 16 hours to 8 hours per day, what happens to the concentration of Na\(^+\)K\(^+\) ATP-ase in its brain?

**Subjects:** Kelly’s subjects were eight golden hamsters.

**Treatments:** There were two treatments: being raised in long (16-hour) days or short (8-hour) days.

**Random assignment of treatments:** To make her study a true experiment, Kelly randomly assigned a day length of 16 hours or 8 hours to each hamster in such a way that half the hamsters were assigned to be raised under each treatment.

**Replication:** Each treatment was given to four hamsters.

**Response variable:** Because Kelly was interested in whether a difference in the amount of light causes a difference in the enzyme concentration, she chose the enzyme concentration for each hamster as her response variable.

**Results:** The resulting measurements of enzyme concentrations (in milligrams per 100 milliliters) for the eight hamsters were

| Short days: | 12.500 | 11.625 | 18.275 | 13.225 |
| Long days: | 6.625 | 10.375 | 9.900 | 8.800 |
You can imagine Kelly defending her design:

**Kelly:** I claim that the observed difference in enzyme concentrations between the two groups of hamsters is due to the difference in the number of hours of daylight.

**Skeptic:** Wait a minute. As you can see, the concentration varies from one hamster to another. Some just naturally have higher concentrations. If you happened to assign all the high-enzyme hamsters to the group that got short days, you'd get results like the ones you got.

**Kelly:** I agree, and I was concerned about that possibility. In fact, that's precisely why I assigned day lengths to hamsters by using random numbers. The random assignment makes it extremely unlikely that all the high-enzyme hamsters got assigned to the same group. If you have the time, I can show you how to compute the probability.

**Skeptic:** (Hastily) That's okay for now. I'll take your word for it. But maybe you can catch me in Chapter 5.

The characteristics of the plan Kelly used are so important that statisticians try to reserve the word *experiment* for studies like hers that answer a question by comparing the results of treatments assigned to subjects at random.

---

**DISCUSSION**

**Cause and Effect**

D20. Plot Kelly’s results in a dot plot, using different symbols for hamsters raised in short days and those raised in long days. In Chapter 9, you’ll see a formal method for analyzing data like these to decide whether the observed difference between the long- and short-day hamsters is too big to be due only to chance. This method will be similar to the sort of logic introduced in the *Martin v. Westvaco* case in Chapter 1. For now, just offer your best judgment: Do you think the evidence supports a conclusion that the number of daylight hours causes a difference in enzyme concentration?

D21. Kelly has shown that hamsters raised in less daylight have higher enzyme concentrations than hamsters raised with more daylight. In order for Kelly to show that less daylight *causes* an increase in the enzyme concentration, she must convince us that there is no other explanation. Has she done that?

---

**Confounding in Observational Studies**

For sample surveys, selecting the sample at random protects against bias, which otherwise can easily mislead you into jumping to false conclusions. For comparative studies, a threat called *confounding* also can lead to false conclusions. Randomizing, when it is possible, protects against this threat in much the same way it protects against bias in sampling.

In everyday language, *confounded* means “mixed up, confused, at a dead end,” and the meaning in statistics is similar.
Two possible influences on an observed outcome are **confounded** if they are mixed together in a way that makes it impossible to separate their effects on the responses.

Studies that claim to show that review courses increase SAT scores often ignore the important concept of confounding. In one study, students at a large high school were offered an SAT preparation course, and SAT scores of students who completed the course were higher than scores of students who chose not to take the course. The positive effect of the review course was confounded with the fact that the course was taken only by volunteers, who would tend to be more motivated to do well on the SAT. Consequently, you can't tell if the higher scores of those who took the course were due to the course itself or to the higher motivation of the volunteers. As you read the next example, ask yourself which influences are being mixed together. Where's the confounding?

The thymus, a gland in your neck, behaves in a peculiar way. Unlike other organs of the body, it doesn't get larger as you grow—it actually gets smaller. Ignorance of this fact led early 20th-century surgeons to adopt a worthless and dangerous surgical procedure. Imagine yourself in their situation. You know that many infants are dying of what seem to be respiratory obstructions, and in your search for a treatment you begin to do autopsies on infants who die with respiratory symptoms. You've done many autopsies in the past on adults who died of various causes, so you decide to rely on those autopsy results for comparison. What stands out most when you autopsy the infants is that they all have thymus glands that look too big in comparison to their body size. Aha! That must be it! The respiratory problems are caused by an enlarged thymus. It became quite common in the early 1900s for surgeons to treat respiratory problems in children by removing the thymus. In particular, in 1912, Dr. Charles Mayo (one of the two brothers for whom the Mayo Clinic is named) published an article recommending removal of the thymus. He made this recommendation even though a third of the children who were operated on died. Looking back at the study, it's easy to spot the confounded variables.

<table>
<thead>
<tr>
<th>Age</th>
<th>Thymus Size</th>
<th>Problems</th>
<th>No evidence</th>
<th>No problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child</td>
<td>Large</td>
<td>No evidence</td>
<td>No problems</td>
<td></td>
</tr>
<tr>
<td>Adult</td>
<td>Small</td>
<td>No evidence</td>
<td>No problems</td>
<td></td>
</tr>
</tbody>
</table>

The doctors couldn't know whether children with a large thymus tend to have more respiratory problems, because they have no evidence about children with a smaller thymus. Age and size of thymus were confounded.

The thymus study is an example of an **observational study**, not an experiment.
Observational studies are less desirable than experiments, but often they are the only game in town. Suppose you want to study accident rates at a dangerous intersection to see if the rate on rainy days is higher than the rate on dry days. It turns out that most of the rainy days you have available for study are weekends and almost all the dry days are during the week. You cannot assign rain to a weekday! The best you can do is design your observational study to cover a long enough period so that you are likely to have some weekdays and weekends in both groups. Otherwise, you have a serious confounding problem.

Don't make the mistake of thinking that observational studies are mere haphazard collections and observations of existing data. Good observational studies, like good experiments, can be designed to help answer specific questions. After researchers take into account every alternative explanation that they can think of, interesting associations can be observed. For example, epidemiologists studying the causes of cancer noticed that lack of exercise seemed to be associated with getting cancer. The best test of a cause-and-effect relationship between amount of exercise and getting cancer would be an experiment in which thousands of people would be assigned randomly to exercise or to not exercise. Researchers would have liked to conduct this experiment, but the practical difficulties were prohibitive in cost and in getting people to comply over the many years the experiment would have to run. To examine this issue further, one research study selected a group of patients with cancer and then matched them according to as many background variables as they could reasonably measure—except how sedentary their jobs had been—with patients who did not have cancer. Those with cancer turned out to have more sedentary jobs. But cause and effect can not be established based on such an observational study alone. The researchers are worried that a confounding variable is present—that people who have sedentary jobs are different in some other way from those who don't. As one researcher said, “The problem is the things we're not smart enough to know about, the things we haven't even thought of.” [Source: Gina Kolata, “But Will It Stop Cancer?”, New York Times, November 1, 2005.]

For drawing conclusions about cause and effect, a good randomized experiment is nearly always better than a good observational study.

These days, any new medical treatment must prove its value in a clinical trial—a randomized comparative experiment. Patients who agree to be part of the study know that a chance method will be used to decide whether they get the standard treatment or the experimental treatment. If Dr. Mayo had used a randomized experiment to evaluate surgical removal of the thymus, he would have seen that the treatment was not effective and many lives might have been spared. However, at the time, randomized experiments were not often used in the medical profession.

**DISCUSSION**

**Confounding in Observational Studies**

D22. Suppose the surgeons had examined infants without respiratory problems and found that their thymus generally was small.

a. Have they now proved that a large thymus causes respiratory problems in children? If so, why? If not, what is another possible explanation?

b. Design an experiment to determine whether removal of the thymus helps children with respiratory problems.
D23. What variables might be confounded with amount of exercise in the observational studies on the association of lack of exercise with cancer?

D24. What is the main difference between an experiment and an observational study?

Factors and Levels

Suppose Kelly decides to add two different diets to her hamster experiment—call them light and heavy. Now her experiment has two factors, diet and length of day, as shown in the table. Each factor has two levels: long and short for length of day, and heavy and light for diet. Choosing one level for length of day and one level for diet gives four possible treatments.

<table>
<thead>
<tr>
<th>Factor 1</th>
<th>Type of Diet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor 2</td>
<td>Light</td>
</tr>
<tr>
<td>Length of Day</td>
<td>Light–Short</td>
</tr>
</tbody>
</table>

In an observational study, nothing is “treated,” but the factor and level terminology still works fine. For example, in the thymus study one factor is the size of the thymus (large or small) and a second factor is age (child or adult).

The Importance of Randomizing

Confounding is the main threat to making reliable inferences about cause. You can think of confounding as the two groups you want to compare differing in some important way other than just on the conditions you want to compare. How can you guard against confounding? The best solution, whenever possible, is to randomize: Use a chance device to decide which people, animals, or objects to assign to each set of conditions you want to compare. If you don’t randomize, it’s risky to generalize.

Wait a minute! Are statisticians asking you to believe that if, say, you want to determine if smoking causes cancer, you should assign people to smoke or not on the basis of a coin flip? Of course not. But it is true that if you can’t use a chance device to assign the conditions you want to compare, for example, smoker and nonsmoker, it becomes very difficult to draw sound conclusions about cause and effect.

Often, you can assign treatments at random to available subjects. For example, to evaluate a special course to prepare students for the SAT, you can start with a group of students who want to take the course. Randomly assign half to take the course and the other half to a group that will not get any special preparation. Then, after both groups of students have taken the SAT, compare their scores. This time, if the students who took the course score 40 points better on average, you can be more confident that the difference is due to the course rather than to some confounding trait of the students willing to take a review course.

Randomizing can protect against confounding.
Why Randomization Makes Inference Possible

- If you assign treatments to units at random, then there are only two possible causes for a difference in the responses to the treatments: chance or the treatments.
- If the probability is small that chance alone will give you such a difference in the responses, then you can infer that the cause of the difference was the treatment.

Your goal in Activity 4.3a is to gain practice with randomization as it is used in the design of experiments and to see the effects of randomization.

**ACTIVITY 4.3a Randomization and Its Effect**

**What you’ll need:** a box, equal-size slips of paper for a random drawing, a coin

The students in your class are to be the subjects in an experiment with two treatments, A and B. The task is to find the best of three ways to assign the treatments to the subjects.

1. Choose two leaders from the class, one for Treatment A and one for Treatment B. The leaders should flip a coin to decide who goes first, then alternately choose class members for their teams (much like choosing sides for a softball game).
2. For each treatment group, record in a table the number of subjects, the proportion of females, the proportion who have brothers or sisters, and the proportion who like to read novels in their spare time. Do the two groups look quite similar or quite different?
3. Next, divide the class by writing the names on the pieces of paper and putting them in a box. Randomly draw out half of them (one by one) to be assigned Treatment A. The names remaining are assigned Treatment B.
4. Repeat step 2 for these groups.
5. Finally, divide the class by having each person flip a coin. Those getting heads are assigned Treatment A, and those getting tails are assigned Treatment B.
6. Repeat step 2 for these new groups.
7. What are the strengths and weaknesses of each method of assigning treatments to subjects? Which method is least random?

A Control or Comparison Group Is Vital

An article on homeopathic medicine in *Time* (“Is Homeopathy Good Medicine?” September 25, 1995, pp. 47–48) began with an anecdote about a woman who had been having pain in her abdomen. She was told to take calcium carbonate, and after two weeks
her pain had disappeared. It was reported that her family “now turns first” to homeopathic medicine. This kind of personal evidence—“It worked for me”—is called anecdotal evidence.

Anecdotal evidence can be useful in deciding what treatments might be helpful and so should be tested in an experiment. However, anecdotal evidence cannot prove, for example, that calcium carbonate causes abdominal pain to disappear. Why not? After all, the pain did go away.

The problem is that pain often tends to go away anyway, especially when a person thinks he or she is receiving good care. When people believe they are receiving special treatment, they tend to do better. In medicine, this is called the placebo effect.

A placebo is a fake treatment, something that looks like a treatment to the patient but actually contains no medicine. Carefully conducted studies show that a large percentage of people who get placebos, but don’t know it, report that their symptoms have improved. The percentage depends on the patient’s problem but typically is over 30%. For example, when people are told they are being treated for their pain, even if they are receiving a placebo, changes in their brain cause natural painkilling endorphins to be released. Consequently, if people given a treatment get better, it might be because of the treatment, because time has passed, or simply because they are being treated by someone or something they trust.

How can scientists determine if a new medication is effective or whether the improvement is due either to the placebo effect or to the fact that many problems get better over time? They use a group of subjects who provide a standard for comparison. The group used for comparison usually is called a control group if the subjects receive a placebo and a comparison group if the subjects receive a standard treatment.

Patients in the treatment group are given the drug to be evaluated. The patients given a placebo get a nontreatment carefully designed to be as much like the actual treatment as possible. The control and treatment groups should be handled exactly alike except for the treatment itself. If a new treatment is to be compared with a standard treatment, the comparison group receives the standard treatment rather than a placebo.

In order for the control or comparison group and the treatment group to be treated exactly alike, both the subjects and their doctors should be “blind.” That is,
the patients should not know which treatment they are receiving, and the doctors who evaluate how much the patients’ symptoms are relieved should be blind as to what the patients received. If only the patients don’t know, the experiment is said to be **blind**. If both the patients and the doctors don’t know, the experiment is said to be **double-blind**.

**Sample Medical Study Designs**

*Comparing a treatment and a placebo.* To test the stimulus effect of caffeine, 30 male college students were randomly assigned to one of three groups of ten students each. Each group was given one of three doses of caffeine (0, 100, and 200 mg), and 2 hours later the students were given a finger-tapping exercise. The measured response was the number of taps per minute. The 0-mg dose can be considered a placebo. Ideally, the three groups would have been given pills that look exactly alike so that they could not tell which treatment they were receiving. This design allows comparisons to be made among two active treatments and a placebo to sort out the effect of caffeine. [Source: Draper and Smith, *Applied Regression Analysis* (John Wiley and Sons, 1981).]

*Comparing a new treatment and a standard treatment.* In an early study on AIDS-related complex (ARC), the drug zidovudine (commonly known as AZT) combined with acyclovir (ACV) was compared to AZT alone (a standard treatment). A total of 134 patients with ARC agreed to participate in the study, which was designed and run as a double-blind, randomized clinical trial. Each patient was randomly assigned either AZT by itself or a combination of AZT and ACV. One of the outcome measures was how many of the ARC patients developed AIDS during the 1 year of the study. A placebo was not used here—such a “treatment” would be unethical because of the health profession’s obligation to administer a treatment thought to be effective. You will analyze the outcome of this experiment in Chapter 8. [Source: David A. Cooper et al., “The Efficacy and Safety of Zidovudine Alone or as Cotherapy with Acyclovir for the Treatment of Patients with AIDS and AIDS-Related Complex: A Double Blind, Randomized Trial,” *AIDS* 7 (1993): 197–207.]

*Comparing a new treatment, a standard treatment, and a placebo.* Another experiment compared an undesirable side effect of some antihistamines—drowsiness—for two treatments (meclastine and promethazine) and a placebo. At the time of the study, meclastine was a new drug and promethazine was a standard drug known to cause drowsiness. Each of nine subjects was given each of the three treatments on different days, in random order. The subjects were blind as to which treatment they were receiving. The outcome measure was the *flicker frequency* of patients (the number of flicks of the eyelids per minute). Low flicker frequency is related to drowsiness, because the eyes are staying shut too long. Note that each subject received all treatments; the only randomization was in the order in which he or she received them. You will analyze the results of this experiment in Chapter 9. [Source: D. J. Hand et al., *A Handbook of Small Data Sets* (London: Chapman and Hall, 1994), p. 8.]

To summarize, a good experiment must have both a random assignment of treatments to units and a control or comparison group that is compared to the group getting the treatment of interest. Such an experiment is called a **randomized comparative experiment**.
A Control or Comparison Group Is Vital

D25. Why is blinding or double-blinding important in an experiment?

D26. A report about a new study to test the effectiveness of the herb St. John’s wort to treat depression says, “The subjects will receive St. John’s wort, an antidepressant drug, or a placebo for at least eight weeks and as long as six months.” Describe how you would design this study to compare the effects of the three treatments. [Source: Los Angeles Times, August 31, 1998, p. A10.]

D27. How would you design an experiment to verify that a placebo effect exists for people who think they are being treated for pain?

Experimental Units and Replication

In educational research, children appear in a group—their classroom—so in order to compare two methods of teaching reading, say, a researcher must assign entire classrooms randomly to the two methods. In such cases, the classroom, not an individual student, is the experimental unit.

Suppose a researcher had six classrooms, each with 25 students, available for her study and assigned three classrooms to each method. The researcher might like to say that each method was replicated on 75 students because the reading ability of 75 students was measured, but, alas, she can claim only that each method was replicated on three classrooms. The students within each classroom were selected as a group and treated in a group setting and so do not contribute independent responses.

**Experimental units** are the people, animals, families, classrooms, and so on to which treatments are randomly assigned. **Replication** is the random assignment of the same treatment to different units.

Performing many replications in an experiment is the equivalent of having a large sample size in a survey—*the more good observations you have, the more faith you have in your conclusions.* (Be careful, however, when someone uses the word replication. Sometimes it is used to mean that your entire experiment was repeated by someone else who came to the same conclusion you did.)

This box summarizes what you have learned about the characteristics of a well-designed experiment.

**Characteristics of a Well-Designed Experiment**

- **Compare.** A treatment group is compared to a control group, or two or more treatment groups are compared to each other.
- **Randomize.** Treatments are randomly assigned to the available experimental units.

(continued)
• Replicate. Each treatment is randomized to enough experimental units to provide adequate assessment of how much the responses from the same treatment vary.

A Tale of Two Sponges

A true story will show why it is critically important to get the units right when you plan a study. Although this study happens to be observational, the lesson about units applies to experiments as well.

It could have been the best of designs, but it was the worst of designs. It looked like the biggest of samples, but in the end it turned out to be the smallest of samples.

Once upon a time, many years ago, there lived a statistically innocent graduate student of biology who wanted to compare the lengths of cells in green sponges and white sponges. After consulting his advisor—who should have known better—the graduate student went to work. For hours at a time, he sat hunched over his microscope, painstakingly measuring the lengths of tiny cells. Weeks went by, and eventually, after he had measured cell lengths for about 700 cells per sponge, the graduate student decided he had enough data. It was time, at last, to do the analysis. This dialogue between the Innocent Graduate Student (IGS) and the project statistician, the Bearer of Bad News (BBN), reveals the tragic ending.

IGS: Here are my pages and pages of numbers. Can I conclude that cells from white and green sponges have different lengths?

BBN: That depends. We’ll have to look at three things—the size of the difference in the white and green averages, the size of your sample, and the size of the natural variability from one unit to the next.

IGS: No problem. My sample size is humongous: 700 cells of each kind. And the variability from one cell to the next isn’t all that big.

BBN: Seven hundred cells certainly is a lot. How many sponges did they come from?

IGS: Uh, just two, one of each color. Does that matter?

BBN: Unfortunately, it’s critical. But I’ll start with the good news. With 700 cells from each sponge, you have rock-solid estimates of...
average cell length for the two sponges you actually looked at. However, . . .

IGS: Right. And because one's green and one's white and the average lengths are different, I can say that . . .

BBN: Not so fast. You have data that let you generalize from the cells you measured to all the cells in your particular two sponges. But that's not the conclusion you were aiming for.

IGS: Yeah. I want to say something about all white sponges compared to all green ones.

BBN: So for you, the kind of variability that matters most is the variability from one sponge to another. Your unit should be a sponge, not a cell.

IGS: You mean to tell me I've got only two units? After all that work!

BBN: I'm afraid you have samples of size 1.

IGS: Woe is me! I'd rather be a character in a bad parody of a Dickens novel. But it will be a far, far better study that I do next time.

The next example describes an experiment with a design that is typical of experiments done in industrial settings.

**Example: Popping Corn**

One Saturday afternoon, you decide to compare the number of kernels left unpopped by a generic and a brand-name popcorn and by your hot air popper and your oil popper. You have the time to pop 20 batches of popcorn before hitting the gym.

a. What are the factors and levels? Describe the treatments.

b. How will you design this study? What are the experimental units? How would randomization be a part of it? Name a variable that might be confounded with your treatments if you didn’t randomize.

c. Have you designed a sample survey, an experiment, or an observational study?

**Solution**

a. The factors are the type of popcorn and the type of popper. There are two levels for type of popcorn: generic and brand-name. There are two levels for type of popper: hot air and oil. The four treatments are the four combinations of the two types of popcorn with the two types of popper. The treatments are represented by the four cells in the table.

<table>
<thead>
<tr>
<th>Type of Popper</th>
<th>Hot Air</th>
<th>Oil</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of Popcorn</td>
<td>Generic</td>
<td>Brand-Name</td>
</tr>
</tbody>
</table>
b. You will have time to pop five batches for each treatment. Because you can’t select bags of popcorn at random (all the bags in your grocery store probably were produced at the same time), you will have to assume—without much justification—that the bags of popcorn are pretty much all alike within each type. You are using the poppers that you have; they aren’t selected at random either.

What you can randomize is the order that you assign to the 20 batches to be popped. You have 20 time periods available to pop batches of popcorn. Your experimental unit is one of these time periods. Place 20 slips of paper in a hat, five with hot air/generic written on them, five with hot air/brand-name, five with oil/generic, and five with oil/brand-name. Draw them out one at a time, popping the next batch according to the treatment. This type of randomization minimizes confounding the treatments with variables such as how hot the poppers get or the changing humidity in your kitchen.

c. The real world is messy and many designed studies cannot be categorized perfectly as a survey, an observational study, or an experiment. This is one of those studies.

• The study described in part b is closest to an experiment. The experimental units are the 20 time periods. Four treatments were to be compared. Each unit had one of the four treatments assigned to it at random. Each treatment was replicated on five time periods.

The random assignment over time reduces confounding with variables related to time, such as the temperature of the room and of the poppers. The statistics alone, however, do not allow generalization of the results beyond the popcorn and poppers used. Such generalizations are made by assuming that the popcorn and the poppers used are representative of the brands or types, but these assumptions cannot be confirmed by the data generated in the experiment. That is why real-world experiments generally are repeated under varying conditions.

• If you didn’t randomize the four treatments to the time periods in your answer to part b, there would be no randomization at all and you would have an observational study.

• If you had been able to use random samples of kernels from each of the two different brands of popcorn, the design would be closer to a sample survey than to an experiment. You would be surveying the two populations of popcorn kernels to determine the percentage that would remain unpopped in each of your two poppers.

Giving a name to the type of study isn’t important. What is important is that you randomize what you can. In the next section, you will learn more about how to deal with variables that cannot be randomly assigned to the experimental units.

### Discussion

#### Experimental Units and Replication

D28. In the teaching-methods study (page 252), why does it seem reasonable that the experimental unit should be the classroom? Provide an example where the experimental unit might be the school.
D29. To study whether women released from the hospital 1 day after childbirth have more problems than women released 2 days after childbirth, many hospitals in a large city were recruited to participate in an experiment. Each participating hospital was randomly assigned to release all women giving routine births either after 1 day or after 2 days of hospital care. An assessment of problems encountered after the women returned home was then done on a random sample of women released from each hospital. What experimental unit should be used as the basis of this study?

D30. Why do you have more faith in your conclusions if your experiment has many replications?

**Summary 4.3: Experiments and Inference About Cause**

The goal of an experiment is to compare the responses to two or more treatments. The key elements in any experiment are

- randomization of treatments to units
- replication of each treatment on a sufficient number of units
- a control or comparison group

Randomly assigning treatments to units allows you to make cause-and-effect statements and protects against confounding. If you can’t randomize the assignment of treatments to units, you have an observational study, and confounding will remain a threat. Confounding makes it impossible to determine whether the treatment or something else caused the response.

The amount of information you get from an experiment depends on the number of replications. Recognizing experimental units is crucial.

So that you have a basis for comparison, experiments require either a control group receiving no treatment or a comparison group receiving a second treatment.

**Practice**

**Cause and Effect**

P14. Research has shown a weak association between living near a major power line and the incidence of leukemia in children. Such a study might measure the incidence of leukemia in children who live near a major power line and compare it to the incidence in children who don’t live near a major power line. Typically, the children in the areas near major power lines are matched by characteristics such as age, sex, and family income to children in the areas not near major power lines.

a. Identify the subjects, conditions, and response variable in such a study.

b. Is this type of study a true experiment? Explain why or why not.

c. According to a newspaper article, “While there is a clear association between high-voltage power lines and childhood leukemia, there is no evidence that the power lines actually cause leukemia.”

[Source: “Power lines tie to cancer unknown,” The Paris [Texas] News, July 30, 2006.] Why might the newspaper come to this conclusion?

P15. Solar thermal systems use heat generated by concentrating and absorbing the sun’s energy to drive a generator and produce electrical power. A manufacturer of solar power generators is interested in comparing the
loss of energy associated with the reflectivity of the glass on the solar panels. (A measure of energy loss is available.) Three types of glass are to be studied. The company has 12 test sites in the Mojave Desert, in the southwestern United States. Each site has one generator. The sites might differ slightly in the amount of sunlight they receive.

a. How would you design an experiment to compare the energy losses associated with reflectivity?

b. What plays the role of “subjects” in your experiment in part a? What is the treatment?

c. How would you change your design if each site had six generators?

P16. Each pair of variables here is strongly associated. Does I cause II, or does II cause I, or is a lurking variable responsible for both?

a. I. wearing a hearing aid or not
   II. dying within the next 10 years or not

b. I. the amount of milk a person drinks
   II. the strength of his or her bones

c. I. the amount of money a person earns
   II. the number of years of schooling

Confounding in Observational Studies

P17. Show the confounding in the SAT study described on page 246 by drawing and labeling a table like the one in the thymus example on page 246.

Factors and Levels

P18. Does the type of lighting or music in a dentist’s waiting room have any effect on the anxiety level of the patient? An experiment to study this question could have nine treatments, represented by this table.

<table>
<thead>
<tr>
<th>Type of Music</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pop</td>
</tr>
<tr>
<td>Low</td>
</tr>
<tr>
<td>Medium</td>
</tr>
<tr>
<td>High</td>
</tr>
</tbody>
</table>

P19. Decades ago, when there was less agreement than there is now about the bad effects of smoking on health, a large study compared death rates of three groups of men—nonsmokers, cigarette smokers, and pipe or cigar smokers. The results are shown here.

<table>
<thead>
<tr>
<th>Deaths (per 1000 men per yr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonsmokers</td>
</tr>
<tr>
<td>Cigarette smokers</td>
</tr>
<tr>
<td>Pipe or cigar smokers</td>
</tr>
</tbody>
</table>

P19. The numbers seem to say that smoking a pipe or cigars almost doubles the death rate, from about 20 to 35 per 1000, but that smoking cigarettes is pretty safe.

a. Do you believe that is true? If not, can you suggest a possible explanation for the pattern?

b. Is this study an observational study or an experiment?

c. What is the factor? What are the levels? What is the response variable?

P20. The investigators also recorded the ages of the men in the study, so it was possible to compare the average ages of the three groups.

<table>
<thead>
<tr>
<th>Average Age (yr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonsmokers</td>
</tr>
<tr>
<td>Cigarette smokers</td>
</tr>
<tr>
<td>Pipe or cigar smokers</td>
</tr>
</tbody>
</table>

Does this information help you account for the pattern of death rates? If so, tell how. If not, tell why not. Now what are the factors in this study?
P21. Driving age. You want to know whether raising the minimum age for getting a driver’s license will save lives. You compare the highway death rates for the 50 U.S. states, grouped according to the legal driving age.

a. Is this study observational or experimental?
b. What is the factor? What are the levels? What is the response variable?
c. Think of a variable that might be confounded with legal driving age.

The Importance of Randomizing

P22. Bears in space. Congratulations! You have just been appointed director of research for Confectionery Ballistics, Inc., a company that specializes in launching gummy bears from a ramp using a launcher made of tongue depressors and rubber bands.

The CEO has asked you to study a variety of factors that are thought to affect launch distance. Your first assignment is to study the effects of the color of the bears. Do green bears travel farther?

Display 4.11 shows actual data for bears launched by members of a statistics class at Mount Holyoke College. Each launch team did ten launches, the first set of five launches using red bears and the second set of five launches using green bears. Launch distances are in inches.

a. Which color bear tended to go farther?
b. Your CEO at Confectionery Ballistics, Inc., wants to reward you for discovering the secret of longer launches. Explain why his enthusiasm is premature.
c. Your CEO is adamant: Color is the key to better launches. In vain you argue that color and launch order are confounded. Finally, your CEO issues an executive order: “Prove it. Show me data.” Determined to meet the challenge, you remember the basic rule of statistics: Plot your data first. Discuss what the plots in Display 4.12 show.
d. How would you change the design of the experiment to eliminate confounding?

<table>
<thead>
<tr>
<th>Launch</th>
<th>Team 1</th>
<th>Team 2</th>
<th>Team 3</th>
<th>Team 4</th>
<th>Team 5</th>
<th>Team 6</th>
<th>Median</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15</td>
<td>44</td>
<td>18</td>
<td>13</td>
<td>10</td>
<td>125</td>
<td>16.5</td>
<td>37.5</td>
<td>44.6</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>40</td>
<td>2</td>
<td>39</td>
<td>35</td>
<td>147</td>
<td>37.0</td>
<td>47.8</td>
<td>50.6</td>
</tr>
<tr>
<td></td>
<td>48</td>
<td>51</td>
<td>10</td>
<td>16</td>
<td>24</td>
<td>35</td>
<td>29.5</td>
<td>30.7</td>
<td>16.8</td>
</tr>
<tr>
<td>Red Bears</td>
<td>25</td>
<td>100</td>
<td>41</td>
<td>41</td>
<td>22</td>
<td>81</td>
<td>41.0</td>
<td>51.7</td>
<td>31.7</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>37</td>
<td>88</td>
<td>41</td>
<td>65</td>
<td>125</td>
<td>53.0</td>
<td>62.5</td>
<td>38.9</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>52</td>
<td>46</td>
<td>31</td>
<td>45</td>
<td>27</td>
<td>38.0</td>
<td>38.3</td>
<td>13.4</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>72</td>
<td>41</td>
<td>41</td>
<td>70</td>
<td>187</td>
<td>71.0</td>
<td>81.8</td>
<td>58.9</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>86</td>
<td>53</td>
<td>55</td>
<td>21</td>
<td>84</td>
<td>54.0</td>
<td>53.3</td>
<td>25.5</td>
</tr>
<tr>
<td></td>
<td>31</td>
<td>89</td>
<td>38</td>
<td>14</td>
<td>33</td>
<td>105</td>
<td>35.5</td>
<td>51.7</td>
<td>36.4</td>
</tr>
<tr>
<td></td>
<td>74</td>
<td>120</td>
<td>92</td>
<td>33</td>
<td>37</td>
<td>174</td>
<td>83.0</td>
<td>88.3</td>
<td>53.4</td>
</tr>
<tr>
<td>Green Bears</td>
<td>31</td>
<td>89</td>
<td>38</td>
<td>14</td>
<td>33</td>
<td>105</td>
<td>35.5</td>
<td>51.7</td>
<td>36.4</td>
</tr>
<tr>
<td></td>
<td>74</td>
<td>120</td>
<td>92</td>
<td>33</td>
<td>37</td>
<td>174</td>
<td>83.0</td>
<td>88.3</td>
<td>53.4</td>
</tr>
<tr>
<td>Median</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>22.5</td>
<td>69.0</td>
<td>43.5</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>29.3</td>
<td>72.1</td>
<td>46.0</td>
</tr>
<tr>
<td>SD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>18.4</td>
<td>30.5</td>
<td>31.1</td>
</tr>
</tbody>
</table>
A Control or Comparison Group Is Vital

P23. In Dr. Mayo’s thymus studies, described on page 246, what did he use as a control group? Could the placebo effect have been a factor in any successful surgeries he had? Was it possible for the study to be blind? Was it possible for the doctors who evaluated how well the patients were doing after surgery to be “blind”?

P24. In a study to see if people have a “magnetic sense”—the ability to use Earth’s magnetism to tell direction—students were blindfolded and driven around in a van over winding roads. Then they were asked to point in the direction of home. They were able to do so better than could be explained by chance alone. [Source: “Tests Point Away from 6th Sense,” Los Angeles Times, July 19, 1982.]

a. Are you satisfied with the design of this study? Are there any other possible explanations for the results other than a magnetic sense?

To improve the design, the experimenter placed magnets on the back of some students’ heads, which were supposed to confuse their magnetic sense, and nonmagnetic metal bars on the back of other students’ heads.

b. What is the control group in the new design?

c. One very important thing wasn’t mentioned in the description of the design of this experiment. What is that?

d. Are you satisfied with the new design?

Later it was determined that the magnetic bars tended to stick to the metal wall of the van.

e. Was this a blind experiment? Was this a double-blind experiment?

Experimental Units and Replication

P25. You want to compare two different textbooks for the AP Statistics course. Ten classes with a total of 150 students will take part in the study. To judge the effectiveness of the books, you plan to use the students’ scores on the AP Statistics Exam, taken at the end of the year. The classes are randomly divided into two groups of five classes each. The first group of five classes, with 80 students...
total, uses the first of the two books. The remaining five classes, with 70 students total, use the other book. Identify the treatments and the experimental units in this study, and then identify the effective sample size for comparing treatments.

P26. You have a summer job working in a greenhouse. The manager says that she has discovered a wonderful new product that will help carnations produce larger blooms, and you decide to design an experiment to check it out. What are your experimental units? How will you use randomization and replication? Do you need a control or comparison group?

P27. On another lazy Saturday afternoon, you decide to compare the strength of two brands of paper towels, Brand A and Brand B, each under the conditions wet and dry. You have one roll of each brand on hand and have time to test 20 towels from each roll before hitting the library. One by one, you stretch each towel in an embroidery hoop, either leaving the towel dry or wetting it with a tablespoon of water, and then place the hoop over a bowl. You place a penny in the center of the hoop, wait 3 seconds, and place another penny on top, continuing until the towel breaks. You record the number of pennies on the paper towel before the towel broke.

a. What are the factors and levels? What is the response variable?

b. How should you use randomization in this study? What are the experimental units?

c. Is your study most like a survey, an observational study, or an experiment?

Exercises

E23. A psychologist wants to compare children from 1st, 3rd, and 5th grades to determine the relationship between grade level and how quickly a child can solve word puzzles. Two schools have agreed to participate in the study. Would this be an observational study or an experiment?

E24. An engineer wants to study traffic flow at four busy intersections in a city. She chooses the times to collect data so that they cover morning, afternoon, and evening hours on both weekdays and weekends. Is this an experiment or an observational study?

E25. Buttercups. Some buttercups grow in bright, sunny fields; others grow in woods, where it is both darker and damper. A plant ecologist wanted to know whether buttercups in sunny locations have adapted in a way that makes them less successful in the shade than in the sun. For his study, he dug up ten plants in sunny locations. Five plants were chosen at random from these ten plants and then were replanted in the sun; the remaining five plants were replanted in the shade. At the end of the growing season, he compared the sizes of the plants.

a. Does this study meet the three characteristics (randomization, replication, control or comparison group) of a true experiment?

b. Why did the plant ecologist bother to dig up the plants that were just going to be replanted in the sunny location? Use the word confounded in your answer.
E26. A metallurgist is studying the properties of copper disks produced by sintering (heating powder until it becomes a solid mass). He has two types of powder available and wants to consider three sintering temperatures. One oven is available for the sintering, and many disks can be made in each run. However, each run can involve only one type of powder.

a. What are the factors and levels? How many treatments are there?
b. How should the metallurgist incorporate randomization into the design of his study?
c. How could you make this study blind?

E27. The college health service at a small residential college wants to see whether putting antibacterial soap in the dormitory bathrooms will reduce the number of visits to the infirmary. In all, 1800 students from 20 dormitories participate. Half the dormitories, chosen at random, are supplied with the special soap; the remaining ones are supplied with regular soap. At the end of one semester, the two groups of students are compared based on the average number of visits to the infirmary per person per semester.

a. What are the units?
b. How many units are there?
c. Is this an observational study or an experiment?

E28. If you’ve studied chemistry or biology, you may know that different kinds of sugar have different molecular structures. Simple sugars, like glucose and fructose, have six carbon atoms per molecule; sucrose, a compound sugar, has twice as many atoms. A biologist wants to know whether complex sugars can sustain life longer than simple sugars. She prepares eight petri dishes, each containing ten potato leafhoppers. Two dishes are assigned to a control group (no food) and two each to a diet of glucose, fructose, and sucrose. The response variable is the time it takes for half the leafhoppers in a dish to die.

a. What are the units?
b. How many units are there?
c. Is this an observational study or an experiment?

E29. Climate and health. Does living in a colder climate make you healthier? To study the effects of climate on health, imagine comparing the death rates in two states with very different climates, such as Florida and Alaska. (The death rate for a given year tells how many people out of every 100,000 died during that year.) In 2000, Florida's death rate was more than twice as high as Alaska's. Why is death rate used as a response variable rather than number of deaths? Why might even death rate not be a good choice for a response variable? (Think about the kinds of people who move to the two states.) Is this an experiment or an observational study?


E30. You are to direct a study to compare the graduation rates of the 35 high schools in your county.

a. How will you define “graduation rate”? Why is it important to look at rates rather than the number of graduates?
b. What issues will determine whether you do a census, a sample survey, an observational study, or an experiment?
c. Will you be able to determine the cause of low graduation rates from the data collected in your study?
Designing Experiments to Reduce Variability

In the previous section, you learned the basics of a good experiment: Randomly assign one of two or more treatments to each experimental unit, and then handle them as alike as possible except for the treatment itself. Within this framework, however, you have further choices about experimental design. If you can protect yourself against confounding by randomizing, designing a good experiment becomes mainly a matter of managing variability. In the first part of this section, you will learn about two types of variability in experiments: one type that you want because it reveals differences between treatments, and another type that you don’t want because it obscures differences between treatments. The rest of the section will show you how a good experimental design can reduce the “bad” kind of variability.

Differences Between Treatments Versus Variability Within Treatments

The results for Kelly’s experiment with hamsters raised in short days and hamsters raised in long days (page 244) are given again in Display 4.13. There is a difference in the response variable between the two treatments: The average enzyme concentration for long days is 8.925, and the average concentration for short days is 13.906, a difference of 4.981. There also is variability within each treatment—like all living things, hamsters will vary, even when treated exactly alike. In fact, for short days, the difference between the largest and smallest enzyme concentrations (18.275 — 11.625, or 6.65) is even larger than the difference between the two treatment means. Still, because the concentration.

E31. Suppose you have 50 subjects and want to assign 25 subjects Treatment A and 25 Treatment B. You flip a coin. If it is heads, the first subject goes into Treatment A; if it is tails, the first subject goes into Treatment B. You continue flipping the coin and assigning subjects until you have 25 subjects in one of the treatment groups. Then the rest of the subjects go into the other treatment group. Does this method randomly divide the 50 subjects into the two treatment groups? Explain.

E32. Suppose you have recruited 50 subjects and want to assign 25 subjects to Treatment A and 25 to Treatment B. For the first subject to arrive on the day of the study, you flip a coin. If it is heads, the first subject goes into Treatment A; if it is tails, the first subject goes into Treatment B. The second subject to arrive then goes into the other treatment group. For the third subject you flip the coin and again make the random assignment. The fourth subject goes into the opposite treatment from that of the third, and so on. (You flip the coin for the odd-numbered subjects only.) Does this method randomly divide the 50 subjects into the two treatment groups? What good property does this method have?
for each of the short days is larger than the concentration for each of the long
days, you probably believe the treatment made a difference.

<table>
<thead>
<tr>
<th>Short Days</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Long Days</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Display 4.13  Dot plots of enzyme concentration in Kelly's
hamsters, in milligrams per 100 milliliters.

Suppose Kelly had gotten results with more spread in the values but the same
means. As Display 4.14 shows, with the values more spread out, it's no longer
obvious that the treatment matters.

<table>
<thead>
<tr>
<th>Short Days</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Long Days</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Display 4.14  Dot plots of altered hamster data.

*In order to conclude that the treatments make a difference, the difference
between the treatments has to be large enough to overshadow the variation within
each treatment.*

A good experimental plan can reduce within-treatment variability and will
allow you to measure the size of the variability that remains both between and
within treatments.

**A Design for Every Purpose**

Now that you've seen some of the general ideas behind designing sound
experiments, it's time to put them into practice. In Activity 4.4a, you will serve
as an experimental unit (subject) for three different versions of an experiment.
The purpose of the experiment is to see whether there is a detectable difference
in heart rate measured under two treatments:

- Treatment 1: standing, with eyes open
- Treatment 2: sitting, with eyes closed, relaxing

The goal of the activity as a whole is for you to think about the advantages and
disadvantages of three different versions of the experiment. If you don't do the
activity, read through it so you will understand the three designs.
Sit or Stand

You will take your pulse several times during this activity, as directed here.

a. Get ready, sitting or standing.

b. Have your teacher time you for 30 seconds. When your teacher says “Go,” start counting beats until your teacher says “Stop.”

c. Double your count to get your heart rate in beats per minute.

Part A: Completely Randomized Design

In Part A, you and your classmates will be assigned randomly to one treatment or the other.

1. Random assignment. Your teacher will pass around a box with slips of paper in it. Half say “stand” and half say “sit.” When the box comes to you, mix up the slips and draw one. Depending on the instruction you get, either stand with your eyes open, or sit with your eyes closed and relax.


3. Record data. Record your heart rate and that of the other students in your treatment group.

4. Summaries. Display the data using side-by-side stemplots for the two treatments. Then compute the mean and standard deviation for each group. Do you think the treatment makes a difference?

Part B: Randomized Paired Comparison Design (Matched Pairs)

In Part B, you and your classmates will first be sorted into pairs based on an initial measurement. Then, within each pair, one person will be randomly chosen to stand and the other will sit.

1. Initial measurement. Take your pulse sitting with your eyes open.

2. Forming matched pairs. Line up in order, from fastest heart rate to slowest; pair off, with the two fastest in a pair, the next two fastest in a pair, and so on.

3. Random assignment within pairs. Either you or your partner should prepare two slips of paper, one that says “sit” and one that says “stand.” One of you should then mix the two slips and let the other person choose one. Thus, within each pair, one of you randomly ends up sitting and the other ends up standing.


5. Record data. Calculate the difference standing minus sitting for each pair of students.

(continued)
6. **Summaries.** Display the set of differences in a stemplot. Then compute the mean and standard deviation of the differences. What should the mean be if the treatment makes no difference? Do you think the observed difference is real or simply due to variation among individuals?

**Part C: Randomized Paired Comparison Design (Repeated Measures)**

This time each person is his or her own matched pair. Each of you will take your pulse under both treatments: standing and sitting. You’ll flip a coin to decide the order.

1. **Random assignment.** Flip a coin. If it lands heads, you will sit first and then stand. If it lands tails, you will stand first.
2. **First measurement.** Take your pulse in the position chosen by your coin flip.
3. **Second measurement.** Take your pulse in the other position.
4. **Record data.** Record your heart rates in a table with the heart rates of other students. Calculate the difference standing minus sitting for each student.
5. **Summaries.** Display the set of differences in a stemplot. Then compute the mean and standard deviation of the differences. Do you think the treatment makes a difference?

### A Design for Every Purpose

If you did not do Activity 4.4a, look at the sample results on pages 266–268 before discussing these questions.

**D31.** Which design do you think is best for studying the effect of position on heart rate: the completely randomized design, the randomized paired comparison design (matched pairs), or the randomized paired comparison design (repeated measures)? Explain what makes your choice better than the other two designs.

**D32.** Which of the three designs do you think is least suitable? Explain what makes this design less effective than the other two. Make up and describe a new scenario (choose a response and two treatments to compare) for which the least suitable design would be more suitable than the other two.

**D33.** Describe the variation within treatments for the design in Part A. How did the other two designs reduce variation within treatments?

### The Completely Randomized Design

In Activity 4.4a, the design in Part A is an example of what statisticians call a completely randomized design. The treatments (sit or stand) are assigned to units completely at random, that is, without any prior sorting or restrictions. For a completely randomized design (CRD), use a chance device to randomly assign a treatment to each experimental unit.
Steps in Creating a Completely Randomized Design (CRD)

1. Number the available experimental units from 1 to $n$.
2. If you have three treatments, for example, use a random number table or your calculator to pick $n/3$ integers at random from 1 to $n$, discarding any repetitions. The units with those numbers will be given the first treatment. Again pick $n/3$ integers, discarding any repetitions. The units with those numbers will be given the second treatment. The remaining units will get the third treatment.

Kelly Acampora’s hamster experiment used a completely randomized design. Each hamster was randomly assigned a treatment. Kelly did this in such a way that her design was balanced; that is, each treatment was assigned to the same number of units.

Sample Results for Sit or Stand: Completely Randomized Design

Display 4.15 provides data for the completely randomized design of Activity 4.4a from one class of 22 students. Each student was randomly assigned to sit or to stand. The parallel dot plots show that the pulses for the “stand” group might be a little higher on average, but the large within-treatment variability obscures a clear view of any real differences between the treatment means.

<table>
<thead>
<tr>
<th>Standing Pulse Rate</th>
<th>Sitting Pulse Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>68</td>
</tr>
<tr>
<td>62</td>
<td>68</td>
</tr>
<tr>
<td>64</td>
<td>60</td>
</tr>
<tr>
<td>88</td>
<td>68</td>
</tr>
<tr>
<td>86</td>
<td>60</td>
</tr>
<tr>
<td>64</td>
<td>62</td>
</tr>
<tr>
<td>72</td>
<td>80</td>
</tr>
<tr>
<td>72</td>
<td>58</td>
</tr>
<tr>
<td>70</td>
<td>60</td>
</tr>
<tr>
<td>48</td>
<td>68</td>
</tr>
<tr>
<td>92</td>
<td>74</td>
</tr>
</tbody>
</table>

Display 4.15  Pulse rates for a completely randomized design.

The Randomized Paired Comparison Design

Matched pairs and repeated measures, the designs in Parts B and C in Activity 4.4a, are examples of randomized paired comparison designs.
In a randomized paired comparison design, pairs of similar units are randomly assigned different treatments.

**Steps in Creating a Randomized Paired Comparison (Matched Pairs) Design**

1. Sort your available experimental units into pairs of similar units. The two units in each pair should be enough alike that you expect them to have a similar response to any treatment.

2. Randomly decide which unit in each pair is assigned which treatment. For example, you could flip a coin, with heads meaning the first unit gets the first treatment and tails meaning the first unit gets the second treatment. The other unit in the pair gets the other treatment.

In the matched pairs design of Part B, the units were sorted into pairs of students having similar sitting heart rates, with treatments randomly assigned within each pair. In the repeated measures design of Part C, both treatments were assigned, in random order, to each person (the ultimate in matching). In experiments of this type, the matched units or the individual units that receive all treatments in random order are called **blocks**.

**Sample Results for Sit or Stand: Matched Pairs and Repeated Measures Designs**

Display 4.16 provides data for the matched pairs design (Part B) of Activity 4.4a for the same 22 students. Students were paired based on their sitting pulse rate. Each pair then flipped a coin to decide who would sit and who would stand. Because the data are paired, it is appropriate to look at the differences between the standing and sitting pulse rates as measures of increase in pulse rate due to standing. The dot plot of these differences shows that well over half the differences are greater than 0.

<table>
<thead>
<tr>
<th>Pair</th>
<th>Standing Pulse Rate</th>
<th>Sitting Pulse Rate</th>
<th>Difference: stand – sit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>64</td>
<td>42</td>
<td>22</td>
</tr>
<tr>
<td>2</td>
<td>66</td>
<td>52</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>58</td>
<td>64</td>
<td>-6</td>
</tr>
<tr>
<td>4</td>
<td>60</td>
<td>58</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>80</td>
<td>82</td>
<td>-2</td>
</tr>
<tr>
<td>6</td>
<td>80</td>
<td>68</td>
<td>12</td>
</tr>
<tr>
<td>7</td>
<td>66</td>
<td>62</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>62</td>
<td>64</td>
<td>-2</td>
</tr>
<tr>
<td>9</td>
<td>54</td>
<td>64</td>
<td>-10</td>
</tr>
<tr>
<td>10</td>
<td>66</td>
<td>60</td>
<td>6</td>
</tr>
<tr>
<td>11</td>
<td>64</td>
<td>50</td>
<td>14</td>
</tr>
</tbody>
</table>

Display 4.16 Difference in pulse rates for a matched pairs design.

Display 4.17 completes the look at data from Activity 4.4a. In this repeated measures design (Part C), each student in the same class as before received each
treatment. Each student flipped a coin to determine whether to sit first or stand first. The dot plot shows that only one of the differences is negative and that 18 of the 22 differences are greater than 0.

<table>
<thead>
<tr>
<th>Student</th>
<th>Standing Pulse Rate</th>
<th>Sitting Pulse Rate</th>
<th>Difference: stand — sit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>62</td>
<td>58</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>70</td>
<td>60</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>60</td>
<td>56</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>68</td>
<td>62</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>78</td>
<td>68</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>56</td>
<td>56</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>72</td>
<td>66</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>68</td>
<td>66</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>64</td>
<td>58</td>
<td>6</td>
</tr>
<tr>
<td>10</td>
<td>72</td>
<td>62</td>
<td>10</td>
</tr>
<tr>
<td>11</td>
<td>58</td>
<td>54</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Student</th>
<th>Standing Pulse Rate</th>
<th>Sitting Pulse Rate</th>
<th>Difference: stand — sit</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>88</td>
<td>88</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>62</td>
<td>40</td>
<td>22</td>
</tr>
<tr>
<td>14</td>
<td>84</td>
<td>78</td>
<td>6</td>
</tr>
<tr>
<td>15</td>
<td>66</td>
<td>62</td>
<td>4</td>
</tr>
<tr>
<td>16</td>
<td>62</td>
<td>58</td>
<td>4</td>
</tr>
<tr>
<td>17</td>
<td>42</td>
<td>42</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>64</td>
<td>70</td>
<td>−6</td>
</tr>
<tr>
<td>19</td>
<td>66</td>
<td>64</td>
<td>2</td>
</tr>
<tr>
<td>20</td>
<td>56</td>
<td>52</td>
<td>4</td>
</tr>
<tr>
<td>21</td>
<td>88</td>
<td>84</td>
<td>4</td>
</tr>
<tr>
<td>22</td>
<td>78</td>
<td>74</td>
<td>4</td>
</tr>
</tbody>
</table>

Display 4.17 Difference in pulse rates for a repeated measures design.

**DISCUSSION**

**Sit or Stand**

D34. Does the randomized paired comparison designs or the completely randomized design provide stronger evidence that the pulse rate tends to be higher when standing than when sitting?

D35. Which of the two randomized paired comparison designs provides stronger evidence that the pulse rate tends to be higher when standing than when sitting?

**The Randomized Block Design**

A randomized paired comparison design is a special type of randomized block design. The difference is that in a generalized block design more than two experimental units may be in each block.
### Steps in Creating a Randomized Block Design

1. Sort your available experimental units into groups (blocks) of similar units. The units in each block should be enough alike that you expect them to have a similar response to any treatment. This is called **blocking**.

2. Randomly assign a treatment to each unit in the first block. (It’s usually best if the same number of units is assigned to each treatment.) Then go to the second block and randomly assign a treatment to each unit in this block. Repeat for each block.

The use of the term *block* in statistics comes from agriculture. One of the earliest published examples of a block design appeared in R. A. Fisher’s *The Design of Experiments* (1935). The goal of the experiment was to compare five types of wheat to see which type gave the highest yield. The five types of wheat were the treatments; the yield, in bushels per acre, was the response; and eight blocks of farmland were available for planting.

There were many possible sources of variability in the blocks of land, mainly differences in the composition of the soil: what nutrients were present, how well the soil held moisture, and so on. Because many of these possibly confounding influences were related to the soil, Fisher made the reasonable assumption that the soil within each block would be more or less uniform and that the variability to be concerned about was the variability from one block to another. His goal in designing the experiment was to keep this between-block variability from being confounded with differences between wheat types.

Fisher’s plan was to divide each of the eight blocks of land “into five plots running from end to end of the block, and lying side by side, making forty plots in all.” These 40 plots were his experimental units. Fisher then used a chance device to assign one type of wheat to each of the five plots in a block. Display 4.18 shows how his plan might have looked. A, B, C, D, and E represent the five types of wheat.

![Display 4.18 An experimental design using blocks.](image)

Suppose it turned out that one of the blocks had really poor (or really favorable) conditions for wheat. Then Fisher’s blocking accomplished two things: All five types of wheat would be affected equally, and the variation within a block could be attributed to the types of wheat.

The effectiveness of blocking depends on how similar the units are in each block and how different the blocks are from each other. Here “similar units” are...
units that would tend to give similar values for the response if they were assigned the same treatments. The more similar the units within a block, the more effective blocking will be.

**Randomized Designs**

D36. What is the main difference between a completely randomized design and a randomized block design?

D37. Why is blocking sometimes a desirable feature of a design? Give an example in which you might want to block and an example in which you might not want to block.

**A Capstone Activity on Experimental Design: Bears in Space**

In this capstone activity, you will return to your job as director of research for Confectionery Ballistics, Inc., the company that specializes in launching gummy bears into space (see P22 on page 258).

Discovering the secret of long launches might not be as important as finding a cure for cancer, but, unlike examples from real science, the bears are something everyone in your class can know firsthand. You're going to return to the bears, but this time color won't matter. For Activity 4.4b, there are two treatments: one book under the ramp and four books under the ramp. The response variable is launch distance. Each team will do a set of ten launches, with teams randomly assigned to use either a steeper or a flatter launch angle. To judge the effects of the treatments, it is important to keep all other potential influences on the response as nearly constant as possible. One important way to do this is to write out a careful protocol that tells you exactly how the experiment is to be done. To write a good protocol, you first have to identify the important sources of within-treatment variation. This requires that you know the situation fairly well, which is one of the main reasons for using the bear-launch study as an example.

**Bears in Space**

**What you'll need:** for each team, a launch ramp and launcher, a coin, a tape measure or yardstick, a supply of gummy bears, four copies of this textbook

1. **Form launch teams and construct your launcher.** Your class should have an even number of teams, at least four. Make your launcher from tongue depressors and rubber bands. First wind a rubber band enough times around one end of a stick to keep it firmly in place. Then place that stick and another together.
and wind a second rubber band tightly around the other end of the two sticks to bind them firmly together. Insert a thin pencil or small wooden dowel between the two sticks to act as a fulcrum.

2. Randomize. Use random numbers or a random ordering of cards containing team names to assign each team to produce data for one or the other of the two treatments:
   - Steeper ramp (four books)
   - Flatter ramp (one book)

   The same number of teams should be assigned each treatment.

3. Organize your team. Decide who will do which job or jobs: hold the launcher on the ramp, load the gummy bear, launch the bear, take measurements, record the data. (When you launch, use a coin instead of your fingers, or you could end up with sore fingers.)

4. Gather data. Each team should do ten launches. After each launch, record the distance. Measure the distance from the front of the ramp, and measure only in the direction parallel to the ramp.

5. Summarize. Plot your team’s ten launch distances on a dot plot. Summarize the spread by computing the standard deviation.

Your plots in Activity 4.4b show variability from three sources: the launch angle, that is, the variability due to the two different treatments (one book, four books); the particular team, that is, the variability from one team to another for teams with the same launch angle; and the individual launches, that is, the variability between launches for the same team. The first is the difference you want to see—the difference between the two treatments. The second two are “nuisance” sources of variability for launches with the same angle; these are to be minimized. In the discussion questions, you will examine these sources of variation.

### Discussion

**Bears in Space: Within-Team Variability**

D38. Why do the launch distances vary so much for your team? List as many explanations as you can think of, and then order them, starting with the one you think has the most effect.

D39. One important source of variation between launches is the effect of practice. Can you think of an alternative strategy that would reduce or come close to eliminating the variation due to practice?
D40. Use your answers to D38 and D39 to write an experimental protocol—a set of rules and steps to follow in order to keep launch conditions as nearly constant as possible. Your protocol should include rules for deciding when, if ever, you should not count a “bad” launch. After each team has written a protocol, your class should discuss them and combine them into a single protocol to be used for future launches.

**DISCUSSION**

**Bears in Space: Between-Team Variability**

D41. Which measure of center—mean or median—gives a better summary of a team’s set of ten launches? Or are they about equally good? Give reasons for your judgment.

D42. Record the team summaries (means or medians) in parallel dot plots, one for each launch angle. Compute the standard deviation of the means (or medians) for each launch angle. Is there more variation between teams with the same launch angle or between launches for the same team (calculated in Activity 4.4b, step 5)? Why do you think the team summaries vary as much as they do?

List as many explanations as you can think of; then order them, starting with the one you think has the most effect. If you were going to summarize a team’s results with a single-number measure of center anyway, why not do only one launch per team and use that number?

D43. What strategies would you recommend for managing the variability between the teams that used the same launch angle?

**DISCUSSION**

**Bears in Space: Difference Between Treatments**

D44. Use the data from the entire class to construct parallel dot plots showing individual launches by team. Think about ways to use the data to measure and compare the sizes of the three kinds of variability—between-treatment, between-team, and within-team. Can you say which variability is largest? Smallest?

D45. Without doing a formal analysis, choose a tentative conclusion from among these three statements, and give reasons for your choice.

I. Launch angle clearly makes a big difference.

II. Launch angle may very well make a difference, but there’s too much variation from other sources to be able to isolate and measure any effect due to launch angle.

III. There’s not much variability anywhere in the data, so it’s safe to conclude that the effect due to launch angle, if any, is quite small.

D46. Combine the data from all the teams that launched with one book into a single group. Do the same for the data from the four-book teams. Make parallel dot plots of one-book launches and four-book launches. Show all individual launches, ignoring teams. Compare the centers and spreads of the two sets of launches. Why might this analysis be misleading?
More on Blocking

In Activity 4.4b, you found a lot of variability, even for the same launch angle.
- Some teams are more skillful than others in launching the bear.
- Within a team, some launches go more smoothly than others.

These two kinds of variability within the treatment can obscure a difference caused by the launch angle.

Ordinarily, when teams carry out the bear-launch experiment, their results show that the variability between teams is so large that it is hard to tell how much difference the angle of the launch ramp makes. Following a careful protocol can reduce the variability somewhat, but there will still tend to be big differences from one team to another. One solution to the problem is to increase the number of experimental units, that is, use more teams.

Although increasing the number of units is a reliable way to get more information about the difference due to treatments, it can be costly or time-consuming. As you’ll see in Activity 4.4c, sometimes there are much better ways to manage variability.

ACTIVITY 4.4c

Block Those Bears!

What you’ll need: the same equipment as in Activity 4.4b

1. **Form teams and assign jobs**, as in Activity 4.4b.
2. **Randomize**. Each team will conduct two sets of five launches, one set with four books under the ramp and one set with just one book. Each team should flip a coin to decide the order of the treatments.
   - Heads: four books first, then one book
   - Tails: one book first, then four books
3. **Gather data**. Carry out the launches, following the protocol your class developed in D40, and record the launch distances.
4. **Compute averages**. List your team’s averages for the two treatments and their difference in a table. Pool the results from all the teams. Based on the differences between the averages, do you think it is reasonable to conclude that launch angle makes a difference, or is it impossible to tell?

DISCUSSION

Block Those Bears!

D47. What kind of experimental design is used in Activity 4.4c? What are the blocks?

D48. **Units**. Remember that you randomly assign a treatment to a unit. For the design of Activity 4.4c, which is the unit: a single launch, a set of five launches, or a set of ten launches by a team?

D49. **Another block design**. Suppose you were to redo Activity 4.4c. Can you think of a way to improve on the design?


**Summary 4.4: Designing Experiments to Reduce Variability**

Random assignment of treatments to experimental units—the fundamental principle of good design—can be accomplished through two basic plans:

- The completely randomized design is characterized by randomly assigning a treatment to each experimental unit while keeping the number of units given each treatment as equal as possible.
- The randomized block design is characterized by placing similar experimental units into different groups, called blocks, and randomly assigning the treatments to the units within each block.

The reason for the randomized block design is that, in any experiment, responses vary not only because of different treatments (between-treatment variation) but also because different subjects respond differently to the same treatment (within-treatment variation). In a well-designed experiment, the experimenter should minimize the amount of variability within the treatment groups because it can obscure the variability resulting from differences between treatments. By taking the blocks into account in analyzing the data, the effect of the within-treatment variability can be minimized.

You have learned two special cases of block design:

- A randomized paired comparison (matched pairs) design involves randomly assigning two treatments within pairs of similar units, such as to twins or to left and right feet.
- A randomized paired comparison (repeated measures) design involves the assignment of all treatments, in random order, to each unit so that comparisons can be made on the same units. An example is a study in which each patient in a clinic is assigned each of three treatments for asthma, in random order.

---

**Practice**

**Differences Between Treatments Versus Variability Within Treatments**

P28. Review the antibacterial soap experiment in E27.

a. List at least two sources of within-treatment variability.

b. Is the point of randomization to reduce the within-treatment variability or to equalize it between the treatment groups?

---

**A Design for Every Purpose**

P29. To test a new drug for asthma, both the new drug and the standard treatment, in random order, will be administered to each subject in the study.

a. What kind of design is this?

b. An observant statistician cries, “No, no! Use two similar subjects in each pair, randomized to each treatment.” What kind of design is this?

c. Which design do you prefer, and why?
P30. In one design of the heart-rate experiment in Activity 4.4a, each person provided one measurement on each treatment. Which design was that?

P31. For each experiment, describe the within-treatment variability that might obscure any difference between treatments. Then describe an experimental design that includes blocking, and define a response variable.

a. To determine whether studying with the radio on helps or hurts the ability to memorize, there will be two treatments: listening to the radio and not listening to the radio. The subjects available are all seniors in the same school.

b. To determine whether adding MSG to soup makes customers eat more soup, a large restaurant will assign two treatments: adding MSG to the soup and not adding MSG. The subjects available are all customers during one evening.

P32. In the experiment to compare drowsiness caused by two antihistamines (Section 4.3, page 251), a new drug, meclastine, was compared to a standard drug, promethazine, and to a placebo. Each subject was given each treatment. Such a design can be used only because the effect of the antihistamine wears off in a relatively short period of time, allowing more than one treatment to be applied to each subject in the study. The response was the average number of eyelid flicks per minute, because low flicker frequency is related to drowsiness. Results are given in Display 4.19.

a. Do the parallel dot plots show much evidence that one drug is performing better than the other? Is it easy to answer the question from these plots? Explain.

b. Display 4.20 shows the scatterplot for the paired data on Treatments A and C. The line on the scatterplot is the line $y = x$. Does it now look as if Treatment A is better than Treatment C? Is it easier to answer this question from the scatterplot that shows the paired data?

c. How else might you compare the responses for Treatments A and C?

d. What is the design of this study? Explain why this design is better than a completely randomized design.
E33. The SAT discussion from Section 4.3 ended with a revised plan that was truly experimental (see pages 248–249). To which of the two types does that design belong, completely randomized or randomized block? Summarize the design by giving the experimental units, factors, levels, response variable, and blocks (if any).

E34. School academic programs often are evaluated on what are called “gain scores,” the gain in academic achievement (as shown by standardized test scores) over the course of an academic year. Such scores often are reported at the school level, not the student level. That is, each school will receive only the mean score of all its children in a given grade. Suppose ten schools have agreed to participate in a study to evaluate the effect of calculator use in teaching mathematics (extensive use versus limited use).

a. In the first design, five participating schools are randomly assigned one of the two treatments and the other five schools are assigned the other treatment. The gain scores in mathematics for the two treatment groups are to be compared at the end of the year. Is this a completely randomized or a randomized block design? Identify the treatments, experimental units, and blocks (if any).

b. In the second design, the participating schools first are paired according to last year’s gain scores in mathematics. Within each pair of schools, one school is randomly assigned to extensive use of calculators and the other to limited use. The gain scores in mathematics are to be compared at the end of the year. Is this a completely randomized or a randomized block design? If the latter, what type of randomized block design is it? Identify the treatments, experimental units, and blocks (if any).

E35. PKU (phenylketonuria) is a disease in which the shortage of an enzyme prevents the body from producing enough dopamine, a substance that helps transmit nerve impulses. To some extent, you can relieve the symptoms by eating a restricted diet, one low in the amino acid phenylalanine. A study was designed to measure the effect of such a diet on the levels of dopamine in the body. The subjects were ten PKU patients. Each patient was measured twice, once after a week on the low phenylalanine diet and again after a week on a regular diet. Identify blocks (if any), treatments, and experimental units, and describe the design used in this study.

E36. In P15, you considered this problem: Solar thermal systems use heat generated by concentrating and absorbing the sun’s energy to drive a generator and produce electrical power. A manufacturer of solar power generators is interested in comparing the loss of energy associated with the reflectivity of the glass on the solar panels. (A measure of energy loss is available.) Three types of glass are to be studied. The company has 12 test sites in the Mojave Desert, in the southwestern United States. Each site has one generator. The sites might differ slightly in terms of the amount of sunlight they receive.

Now the manufacturer wants to expand the test sites to include generators that are available in the mountains of Tennessee, the desert of northern Africa, and central Europe. Each of these locations has 12 test sites.
sites, with one generator at each site. Further, there are two designs of panels in which to place the glass. Describe how you would design this study, identifying treatments, experimental units, and blocks (if any).

E37. Finger tapping. Many people rely on the caffeine in coffee to get them going in the morning or to keep them going at night. But wouldn’t you rather have chocolate? Chocolate contains theobromine, an alkaloid quite similar to caffeine both in its structure and in its effects on humans. In 1944, C. C. Scott and K. K. Chen reported the results of a study designed to compare caffeine, theobromine, and a placebo. Their design used four subjects and randomly assigned the three treatments to each subject, one treatment on each of three different days. Subjects were trained to tap their fingers in a way that allowed the rate to be measured—presumably, this training eliminated any practice effect. The response was the rate of tapping 2 hours after taking a capsule containing either one of the drugs or the placebo. Are treatments assigned to the experimental units completely at random or using a block design? Identify the response, experimental units, treatments, and blocks (if any). [Source: C. C. Scott and K. K. Chen, “Comparison of the Action of 1-ethyl Theobromine and Caffeine in Animals and Man,” Journal of Pharmacologic Experimental Therapy 82 (1944): 89–97.]

E38. Walking babies. The goal of this experiment was to compare four 7-week training programs for infants to see whether special exercises could speed up the process of learning to walk. One of the four training programs was of particular interest: It involved a daily 12-minute set of special walking and placing exercises. The second program (the exercise control group) involved daily exercise for 12 minutes but without the special exercises. The third and fourth programs involved no regular exercise (parents were given no instructions about exercise) but differed in their follow-up: Infants in the third program were checked every week, whereas those in the fourth program were checked only at the end of the study. Twenty-three 1-week-old babies took part, and each was randomly assigned to one of the groups. The response was the time, in months, when the baby first walked without help.

Tell how the treatments are assigned to the experimental units—completely at random or using a block design. Identify the response, experimental units, treatments, and blocks (if any).

E39. Sawdust for lunch? Twenty-four tobacco hornworms served as subjects for this experiment, which was designed to see how worms raised on low-quality food would compare with worms raised on a normal diet. The two dozen worms were randomly divided into two groups, and the lucky half was raised on regular worm food. The unlucky half got a mixture of 20% regular food and 80% cellulose. Cellulose has no more food value for a hornworm than it has for you—neither you nor a hornworm can digest the stuff. The experimenter kept track of how much each hornworm ate and computed a response value based on the total amount eaten in relation to body weight.

Tell how treatments were assigned to experimental units in this experiment—completely at random or using a block design. Identify the response, experimental units, treatments, and blocks (if any).
E40. Arthritis is painful, and those who suffer from this disease often take pain relief and anti-inflammatory medication for long periods of time. One of the side effects of such medications is that they often cause stomach damage such as lesions and ulcers. The goal of research, then, is to find an anti-inflammatory drug that causes minimal stomach damage. An experiment was designed to test two treatments for arthritis pain (old and new) against a placebo by measuring their effect on lesions in the stomachs of laboratory rats. Rats with similar stomach conditions at the start of the experimental period were randomly assigned to one of the three treatments. Total length of stomach lesions (in millimeters) was measured in each rat after a 2-week treatment period.

a. Was the treatment assignment in this experiment completely at random or in blocks? Does this seem like the best way to make an assignment in this case?

b. What are the experimental units? What is the response measurement?

c. Is the use of a placebo essential here?

Chapter Summary

There are two main types of chance-based methods of data collection: sampling methods and experimental designs. Sampling methods, studied in the first part of this chapter, use chance to choose the individuals to be studied. Typically, you choose individuals in order to ask them questions, as in a Gallup poll; thus, samples and surveys often go together. Experiments, introduced in the second part of this chapter, are comparative studies that use chance to assign the treatments you want to compare. An experiment should have three characteristics: random assignment of treatments to units, two or more treatments to compare, and replication of each treatment on at least two subjects.

The purposes of sampling and experimental design are quite different. Sample surveys are used to estimate the parameters of fixed, well-defined populations. Experiments are used to establish cause and effect by comparing treatments.

Display 4.21 summarizes the differences between a survey and an experiment.

Randomization is the fundamental principle in both types of study because it allows the use of statistical inference (developed in later chapters) to generalize results. In addition, random selection of the sample reduces bias in surveys,
and random assignment of treatments to subjects reduces confounding in experiments.

A careful observational study can suggest associations between variables, but the possibility of a lurking variable being responsible for the association means inferences cannot be made.

Display 4.22 shows how sample surveys, experiments, and observational studies fit together. Any statistical study has at its core the selection of units on which measurements will be made and the possible assignment of conditions (or treatments) to those units. The sample surveys discussed in this chapter fall into the upper-right box, the experiments into the lower-left box, and the observational studies into the lower-right box. Studies of the type described in the upper-left box, which use both random sampling and random assignment of treatments, have not been discussed in this chapter.

<table>
<thead>
<tr>
<th>Selection of Units</th>
<th>Assignment of Conditions or Treatments to Units</th>
<th>Inference to Population Appropriate?</th>
</tr>
</thead>
<tbody>
<tr>
<td>At Random</td>
<td>Experiment with broad scope of inference: A random sample of units is selected from a population; treatments are randomly assigned to the units.</td>
<td>Inferences to the population can be drawn.</td>
</tr>
<tr>
<td>At Random</td>
<td>Sample survey: A random sample of units is selected from a population; the conditions are already built into the units.</td>
<td></td>
</tr>
<tr>
<td>Not at Random</td>
<td>Experiment with narrow scope of inference: A group of available units is used; treatments are randomly assigned to the units.</td>
<td>Inferences are limited to only the units included in the study.</td>
</tr>
<tr>
<td>Not at Random</td>
<td>Observational study: A group of available units is used; the conditions are already built into the units.</td>
<td></td>
</tr>
<tr>
<td>Causal Inference Appropriate?</td>
<td>Causal inferences can be drawn.</td>
<td>Associations may be observed but no causal inferences drawn.</td>
</tr>
</tbody>
</table>


Review Exercises

E41. For each situation, tell whether it is better to take a sample or a census, and give reasons for your answer.

<table>
<thead>
<tr>
<th>Characteristic of Interest</th>
<th>Population of Interest</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Average life of a battery</td>
<td>Alkaline AAA batteries</td>
</tr>
<tr>
<td>b. Average age</td>
<td>Current U.S. senators</td>
</tr>
<tr>
<td>c. Average price per gallon</td>
<td>Purchases of regular-octane gasoline sold at U.S. stations next week</td>
</tr>
</tbody>
</table>

E42. You want to estimate the percentage of people in your area with heart disease who also smoke cigarettes. The people in your area who have heart disease make up your population. You take as your frame all records of patients hospitalized in area hospitals within the last 5 years with a diagnosis of heart disease. How well do you think this frame represents the population? If you think bias is likely, identify what kind of bias it would be and explain how it might arise.
E43. Consider using your statistics class as a convenience sample in each of these situations. For each, tell whether you think the sample will be reasonably representative and, if not, in what way(s) you expect your class to differ from the given population of interest.

<table>
<thead>
<tr>
<th>Characteristic of Interest</th>
<th>Population of Interest</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Percentage who can curl their tongues</td>
<td>U.S., age 12 or older</td>
</tr>
<tr>
<td>b. Average age</td>
<td>U.S., all adults</td>
</tr>
<tr>
<td>c. Average blood pressure</td>
<td>U.S., students your age</td>
</tr>
<tr>
<td>d. Percentage who prefer science to English</td>
<td>U.S., students your age</td>
</tr>
<tr>
<td>e. Average blood pressure</td>
<td>U.S., all adults</td>
</tr>
</tbody>
</table>

E44. The U.S. Bureau of Labor Statistics collects data on occupational variables from a nationwide sample of households. This paragraph is from its news bulletin “Union Members in 1996” [January 31, 1997, p. 3]: “The data also are subject to nonsampling error. For example, information on job-related characteristics of the worker, such as industry, occupation, union membership, and earnings, are sometimes reported by a household member other than the worker. Consequently, such data may reflect reporting error by the respondent. Moreover, in some cases, respondents might erroneously report take-home pay rather than gross earnings, or may round up or down from actual earnings.”

a. Would reporting take-home pay rather than gross earnings increase or decrease the estimate of the earnings of various types of workers?

b. What other sources of nonsampling error might this study contain?

E45. A friend who wants to be in movies is interested in how much actors earn and has decided to gather data using a simple random sample. The World Almanac and Book of Facts has a list of actors that your friend plans to use as a sampling frame. Would you advise against using that list as a frame? What sort of bias do you expect?

E46. What is the average number of representatives per state in the U.S. House of Representatives? If you really want to know, you should use a census rather than a sample, but because you like statistics so much you’ve decided to use a random sample. Tell which of these two sampling methods is biased and describe the bias: Will estimates tend to be too high or too low?

Method I. Start with a list of all the current members of the House. Take a simple random sample of 80 members, and for each representative chosen record the number of representatives from that person’s state. Then take the average.

Method II. Start with a list of all 50 states. Take a simple random sample of 5 states, and for each state chosen record the number of representatives from that state. Then take the average.

E47. For an article on used cars, a writer wants to estimate the cost of repairs for a certain 2000 model of car. He plans to take a random sample of owners who bought that model from a compilation of lists supplied by all U.S. dealers of that make of car. He tells his research assistant to send letters to all the people in the sample, asking them to report their total repair bills for last year. His research assistant tells him, “You might want to think again about your sampling frame. I’m afraid your plan will miss an important group of owners.” What group or groups?

E48. If you look up Shakespeare in just about any book of quotations, you’ll find that the listing goes on for several pages. If you look up less famous writers, you find much shorter listings. What’s the average number of quotations per author in, say, the Oxford Dictionary of Quotations, 3rd edition? Suppose you plan to base your estimate on a random sample taken by this method: You take a simple random sample of pages from the book. For each page chosen, you find the first quotation in the upper-left corner of that page and record the number of quotations by that author. Then you find
the average of these values for all the pages in your sample. What's wrong with this sampling method?

E49. You've just been hired as a research assistant with the state of Maine. For your first assignment, you're asked to get a representative sample of the fish in Moosehead Lake, the state's largest and deepest lake. Your supervisor, who knows no statistics and can't tell a minnow from a muskellunge, tells you to drag a net with 1-in. mesh (hole size) behind a motorboat, up and down the length of the lake. “No,” you tell him. “That method is biased!” What is the bias? What kinds of fish will tend to be overrepresented? Underrepresented? (You don't have to know the difference between a minnow and a muskie to answer this. You do have to know a little statistics.)

E50. You are asked to provide a forestry researcher with a random sample of trees on a 1-acre lot. Your supervisor gives you a map of the lot showing the location of each tree and tells you to choose points at random on the map and, for each point, to take the tree closest to the point for your sample. “Sorry, sir,” you tell him. “That sampling method was once in common use, but then someone discovered it was biased. It would be better to number all the trees in the map and use random numbers to take an SRS.” Draw a small map, with about ten trees, that you could use to convince your supervisor that his method is biased. Put in several younger trees, which grow close together, and two or three older trees, whose large leaf canopies discourage other trees from growing nearby. Then explain which trees are more likely to be chosen by the method of random points and why.

E51. You’re working as an assistant to a psychologist. For subjects in an experiment on learning, she needs a representative sample of 20 adult residents of New York City. She tells you to run an ad in the New York Times, asking for volunteers, and to randomly choose 20 from the list of volunteers. What’s the bias?

E52. Give an example in which nonresponse bias is likely to distort the results of a survey.

E53. Tell how to carry out a four-stage random sample of voters in the United States, with voters grouped by precinct, precincts grouped by congressional district, and congressional districts grouped by state.

E54. Tell how to carry out a multistage sample for estimating the average number of characters per line for the books on a set of shelves.

E55. A researcher designs a study to determine whether young adults who follow a Mediterranean diet (lots of fruits, vegetables, grains, and olive oil; little red meat) end up having fewer heart attacks in middle age than those who follow an average American diet. She randomly selects 1000 young adults in Greece and 1000 young adults in the United States. These 2000 young adults are categorized according to whether they follow a Mediterranean diet or a typical American diet. Thirty years later, she finds that a lower percentage of those who followed the Mediterranean diet have had a heart attack.

a. For the purpose of determining whether a Mediterranean diet results in fewer heart attacks than a typical American diet, is this an observational study or an experiment? Explain.

b. Explain why the researcher cannot conclude that following a Mediterranean diet results in a lower chance of having a heart attack than following an average American diet.

c. What factors might be confounded with diet?

E56. A mother does not keep cola in her house because she is convinced that cola makes her 7-year-old daughter hyperactive. Nevertheless, the daughter gets a cola whenever she has fast food or goes to a birthday party, and the mother has noticed that she then acts hyperactive.

a. Has the mother done an experiment or an observational study?
b. Name two factors that might be confounded with cola, and explain how they are confounded with cola.

c. Design a study that would allow the mother to determine whether giving her daughter cola makes her hyperactive. Are there any potential problems with such a study?

E57. For each science project described,

i. classify the project as a survey, an observational study, or an experiment

ii. if a survey, name the explanatory variables; if an observational study or experiment, name the factors and levels

iii. name the response variable

iv. if a survey, give the sample size; if an observational study, give the number of observed units; if an experiment, give the number of experimental units

a. A student wanted to study the effect of social interaction on weight gain in mice. A large number of baby mice from different mothers were weaned, weighed, and then randomly divided into four groups. One group lived with their mother and siblings, one group lived with their mother but not their siblings, one group lived only with their siblings, and each mouse in the last group lived in isolation. All groups were otherwise treated alike and were given as much food as they wanted. After 90 days, the mice were weighed again.

b. A student picked four different sites on an isolated hillside at random. At each site, he measured off a 10-ft-by-10-ft square. At each site, a sample of soil was taken and the amounts of ten different nutrients were measured. The student counted the number of species of plants at each site, hoping to be able to predict the number of species from the amounts of the nutrients.

c. A college professor helped his daughter with a 2nd-grade science project titled “Does Fruit Float?” They tested 15 different kinds of fruit and classified them as “floaters” and “nonfloaters.” For the display, they took a photograph of a watermelon floating in a swimming pool and a grape sitting at the bottom. The daughter wrote her own explanation of what determines whether a particular type of fruit floats.

E58. You want to test whether your company’s new shampoo protects against dandruff better than the best-selling brand. Unfortunately, the best-selling brand has a very distinctive color, which you are unable to duplicate in your new shampoo. Design a randomized paired comparison experiment to test your shampoo. Discuss any confounding you expect to encounter.

E59. Investigators found seven pairs of identical twins in which one twin lived in the city and the other in the country, and both were willing to participate in a study to determine how quickly their lungs cleared after they inhaled a spray containing radioactive Teflon particles. Display 4.23 gives the percentage of radioactivity remaining 1 hour after inhaling the spray.

<table>
<thead>
<tr>
<th>Twin Pair</th>
<th>Rural</th>
<th>Urban</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10.1</td>
<td>28.1</td>
<td>-18.0</td>
</tr>
<tr>
<td>2</td>
<td>51.8</td>
<td>36.2</td>
<td>15.6</td>
</tr>
<tr>
<td>3</td>
<td>33.5</td>
<td>40.7</td>
<td>-7.2</td>
</tr>
<tr>
<td>4</td>
<td>32.8</td>
<td>38.8</td>
<td>-6</td>
</tr>
<tr>
<td>5</td>
<td>69.0</td>
<td>71.0</td>
<td>-2</td>
</tr>
<tr>
<td>6</td>
<td>38.8</td>
<td>47.0</td>
<td>-8.2</td>
</tr>
<tr>
<td>7</td>
<td>54.6</td>
<td>57.0</td>
<td>-2.4</td>
</tr>
</tbody>
</table>


a. What is the factor? What are its levels? Identify a block.
b. Is this an observational study or an experiment? Explain.

c. Is there more variation in response within the urban twins, within the rural twins, or in the differences? Do you believe the study demonstrates that environment makes a difference in this case?

d. Why were twins used? What kind of variation is reduced by using identical twins?

E60. Design a taste-comparison experiment. Suppose you want to rate three brands of chocolate chip cookies on a scale from 1 (terrible) to 10 (outstanding). Tell how to run this experiment as a randomized block design.

E61. Exercise bikes. You work for a gym and have been asked to design an experiment to decide which of two types of exercise bikes will get the most use. Your exercise bike room has space for eight bikes, which you will place as in Display 4.24. People enter the room through the door at the top of the diagram. A counter on each bike records the number of hours it has been used. Design an experiment to compare the number of hours the two different brands of exercise bike are used that takes into account the fact that some bikes are in locations that make them more likely to be used than others when the gym isn’t full.

E62. Read one of these articles from Statistics: A Guide to the Unknown (Brooks/Cole, 2006), and write a paragraph explaining how randomization and control or comparison groups were used in the study reported in the article.

1. Jennifer Hill, “Evaluating School Choice Programs”


AP1. Researchers want to estimate the mean number of children per family, for all families that have at least one child enrolled in one of ten similarly sized county high schools. Which sampling plan is biased?

A Randomly select 100 families from those families in the county that have at least one child in high school. Compute the mean number of children per family.
B Randomly select one high school, and compute the mean number of children in each family with a child or children in that school.
C Randomly select 20 students from each of the ten high schools, and ask each one how many children are in his or her family. Compute the mean number of children per family.
D From a list of the families in all ten high schools, choose a random starting point and then select every tenth family. Compute the mean number of children per family.
E None of the above is a biased plan.

AP2. A radio program asked listeners to call in and vote on whether the notorious band, The Rolling Parameters, should perform at the Statistics Day celebration. Of the 956 listeners who responded, 701 answered that The Rolling Parameters should perform. Which type of sampling does this example use?

A stratified random
B cluster
C systematic
D quota
E voluntary response

AP3. To select students to explain homework problems, a teacher has students count off by 5’s. She then randomly selects an integer from 1 through 5. Every student who counted off that integer is asked to explain a problem. Which type of sampling plan is this?

A convenience
B systematic
C simple random
D stratified random
E cluster

AP4. To conduct a survey to estimate the mean number of minutes adults spend exercising, researchers stratify by age before randomly selecting their sample. Which of the following is not a good reason for choosing this plan?

A Without stratification by age, age will be confounded with the number of minutes reported.
B Researchers will be sure of getting adults of all ages in the sample.
C Researchers will be able to estimate the mean number of minutes for adults in various age groups.
D Adults of different ages may exercise different amounts on average, so stratification will give a more precise estimate of the mean number of minutes spent exercising.
E All of these are good reasons.

AP5. A movie studio runs an experiment in order to decide which of two previews to use for its advertising campaign for an upcoming movie. One preview features the movie’s romantic scenes and is expected to appeal more to women. The other preview features the movie’s action scenes and is expected to appeal more to men. Sixteen subjects take part in this experiment, eight women and eight men. After viewing one of the previews, each person will rate how much he or she wants to see the movie. Which of the following best describes how blocking should be used in this experiment?

A Use blocking, with the men in one block and the women in the other.
B Use blocking, with half the men and half the women in each block.
C Do not block, because the preview that is chosen will have to be shown to audiences consisting of both men and women.
D Do not block, because the response will be confounded with gender.
E Do not block, because the number of subjects is too small.
AP6. A recent study tried to determine whether brushing or combing hair results in healthier-looking hair. Forty male volunteers were randomly divided into two groups. One group only brushed their hair and the other group only combed their hair. Other than that, the volunteers followed their usual hair care procedures. After two months, an evaluator who did not know the treatments the volunteers used, scored each head of hair by how healthy it looked. There was almost no difference in the scores of the two treatment groups. Which statement best summarizes this study?

- If a male wants healthy looking hair, it probably doesn't matter whether he brushes or combs it.
- You can't tell whether brushing or combing is better, because the treatments are likely to be confounded with variables such as which kind of shampoo a male uses.
- You can't come to a conclusion, because the study wasn't double-blind.
- You can't come to a conclusion, because only volunteers were included.
- The sample size is too small for any conclusion to be drawn.

AP7. Which of the following is not a necessary component of a well-designed experiment?

- There must be a control group that receives a placebo.
- Treatments are randomly assigned to experimental units.
- The response variable is the same for all treatment groups.
- There are a sufficient number of units in each treatment group.
- All units are handled as alike as possible, except for the treatment.

AP8. In a clinical trial, a new drug and a placebo are administered in random order to each subject, with six weeks between the two treatments. Which best describes this design?

- completely randomized with blocking
- completely randomized with no blocking
- randomized paired comparison (matched pairs)
- randomized paired comparison (repeated measures)
- two-stage randomized

### Investigative Tasks

AP9. Needle threading. With your eye firmly fixed on winning a Nobel prize, you decide to make the definitive study of the effect of background color (white, black, green, or red) on the speed of threading a needle with white thread. Design three experiments—one that uses no blocks, one that creates blocks by grouping subjects, and one that creates blocks by reusing subjects. Tell which of the three plans you consider most suitable, and why.

AP10. For the 2000 U.S. Census, controversy erupted over the Census Bureau proposal to use sampling to adjust for the anticipated undercount. Here is a simplified version of the plan: The Census Bureau collects the information mailed in by most of the residents of a region. Some residents, however, did not receive forms or did not return them for some reason; these are the uncounted persons. The bureau now selects a sample of blocks (neighborhoods) from the region and sends field workers to find all residents in the sampled blocks. The residents the field workers found are matched to the census data, and the number of residents uncounted in the original census is noted. The census count for those blocks is then adjusted according to the proportion uncounted. (For example, if one-tenth of the residents were uncounted, the original census figures are adjusted upward by 11%.) In addition, the same adjustment factor is used for neighboring regions that have characteristics similar to the region sampled. Comment on the strengths and weaknesses of this method.
Were some types of people on the Titanic more likely to go down with the ship than others? Basic concepts of probability can help you decide.
You have been using the idea of probability since Chapter 1. In your study of *Martin v. Westvaco*, a question was this: If you select three workers at random for layoff from ten workers ages 25, 33, 35, 38, 48, 55, 55, 55, 56, and 64, what is the probability that the mean of the three ages will be 58 or more? You estimated this probability by repeating the process of randomly drawing three ages and computing the mean age. Because all possible sets of three ages are equally likely, you divided the number of times you got a mean age of 58 or more by the total number of runs in your simulation:

\[
\text{probability} = \frac{\text{number of runs with mean age 58 or more}}{\text{total number of runs}}
\]

You had a model—selecting three workers at random—that let you predict how things are supposed to behave. You also had data—from the court case—that told you how things actually did behave. The key question, then and for the rest of this book, is this: Are the data consistent with the model? Or should you scrap the model and look for some other way to account for the data? As in Chapter 1, the answer will use the language of probability.

Before going any further into statistics, you need to learn more about building and using probability models. In Chapter 1, you used simulation to get an approximate distribution of all possible average ages when three workers are picked at random for layoff. Sometimes, instead of doing a simulation, you can use the rules of probability to construct exact distributions. That’s what you’ll learn to do in this chapter and the next.

**In this chapter, you will learn to**

- list all possible outcomes of a chance process in a systematic way
- design simulations and use them to estimate probabilities
- use the Addition Rule to compute the probability that event $A$ or event $B$ (or both) occurs
- use the Multiplication Rule to compute the probability that event $A$ and event $B$ both occur
- compute conditional probabilities, the probability that event $B$ occurs given that event $A$ occurs

As you work through the chapter, you will see a number of problems involving coins, cards, and dice. These familiar devices effectively demonstrate fundamental principles that apply to many probabilistic situations. As you gain understanding, you’ll be able to apply what you’ve learned to new statistical problems.
Food companies routinely conduct taste tests to make sure their new products taste at least as good as the competing brand. Jack and Jill, co-owners of Downhill Research, just won a contract to determine if people can tell tap water from bottled water. They will give each person in their sample both kinds of water, in random order, and ask which is the tap water.

Jack and Jill know they have a long climb ahead of them, but they’re prepared to take it a step at a time. Before they start, they want to know what might happen if people can’t identify tap water. For example, they want to know: If we ask people which is the tap water and none of the people really know, what is the probability they all will guess correctly?

Jack and Jill know several basics about probability, described here.

### Fundamental Facts About Probability

- **An event** is a set of possible outcomes from a random situation.
- Probability is a number between 0 and 1 (or between 0% and 100%) that tells how likely it is for an event to happen. At one extreme, events that can’t happen have probability 0. At the other extreme, events that are certain to happen have probability 1.
- If the probability that event $A$ happens is denoted $P(A)$, then the probability that event $A$ doesn’t happen is $P(\text{not } A) = 1 - P(A)$. The event $\text{not } A$ is called the complement of event $A$.
- If you have a list of all possible outcomes and all outcomes are equally likely, then the probability of a specific outcome is

\[
\frac{1}{\text{the number of equally likely outcomes}}
\]

and the probability of an event is

\[
\frac{\text{the number of outcomes in the event}}{\text{the number of equally likely outcomes}}
\]

Jack and Jill begin by formulating a model that specifies that if a person can’t identify tap water, then he or she will choose the tap water, $T$, with probability 0.5 and the bottled water, $B$, with probability 0.5. If they have only one taster, that is, $n = 1$, then the probability that the person will guess correctly is written $P(T) = 0.5$.

Onward and upward: What if they have two tasters, that is, $n = 2$? Assuming that the tasters can’t identify tap water, what is the probability that both people will guess correctly and choose $T$? Here the research stumbles.

**Jack:** There are three possible outcomes: Neither person chooses $T$, one chooses $T$, or both choose $T$. These three outcomes are equally
likely, so each outcome has probability \( \frac{1}{3} \). In particular, the probability that both choose \( T \) is \( \frac{1}{3} \).

**Jill:** Jack, did you break your crown already? I say there are four equally likely outcomes: The first taster chooses \( T \) and the second also chooses \( T \) (\( TT \)); the first chooses \( T \) and the second chooses \( B \) (\( TB \)); the first chooses \( B \) and the second chooses \( T \) (\( BT \)); or both choose \( B \) (\( BB \)). Because these four outcomes are equally likely, each has probability \( \frac{1}{4} \). In particular, the probability that both choose \( T \) is \( \frac{1}{4} \), not \( \frac{1}{3} \).

A **probability distribution** gives all possible values resulting from a random process and the probability of each. Display 5.1 summarizes Jill’s probability distribution.

<table>
<thead>
<tr>
<th>Number Who Choose ( T )</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>( \frac{1}{4} )</td>
</tr>
<tr>
<td>1</td>
<td>( \frac{2}{4} )</td>
</tr>
<tr>
<td>2</td>
<td>( \frac{1}{4} )</td>
</tr>
</tbody>
</table>

Display 5.1 Jill’s probability distribution for \( n = 2 \).

**DISCUSSION**

**Probability Distributions**

D1. Copy Jill’s table, leaving out the probabilities. Fill in new probabilities so they represent Jack’s probability distribution.

D2. What is the sum of the probabilities in a probability distribution? Why must this be so?

D3. Who do you think is right, Jack or Jill? Give your reason(s). How could Jack and Jill decide which of them is right?

**Where Do Probabilities Come From?**

To determine which of them has assigned the correct probabilities, Jack and Jill decide to be philosophical and think about how probabilities can be assigned. Probabilities come from three different sources:

- **Observed data** (long-run relative frequencies). For example, observation of thousands of births has shown that about 51% of newborns are boys. You can use these data to say that the probability of the next newborn being a boy is about 0.51.

- **Symmetry** (equally likely outcomes). If you flip a coin and catch it in the air, nothing about the physics of coin flipping suggests that one side is more likely than the other to land facing up. Based on symmetry, it is reasonable to think that heads and tails are equally likely, so the probability of heads is 0.5.

- **Subjective estimates**. What’s the probability that you’ll get an A in this statistics class? That’s a reasonable, everyday kind of question. The use of probability
is meaningful in this situation, but you can't gather data or list equally likely outcomes. However, you can make a subjective judgment.

Based on symmetry, Jack and Jill agree that if \( n = 1 \), then \( P(T) = 0.5 \). They both used symmetry to justify their models for \( n = 2 \), saying that their outcomes were equally likely, but they know that one of them must be wrong. To decide which one, they decide to collect some data. They use two flips of a coin to simulate the taste-test experiment with two tasters who can't identify tap water. Two tails represented neither person choosing the tap water, one heads and one tails represented one person choosing the tap water and the other choosing the bottled water, and two heads represented both people choosing the tap water. Hours later, they have their results, which are given in Display 5.2.

<table>
<thead>
<tr>
<th>Number Who Choose T</th>
<th>Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>782</td>
<td>0.26</td>
</tr>
<tr>
<td>1</td>
<td>1493</td>
<td>0.50</td>
</tr>
<tr>
<td>2</td>
<td>725</td>
<td>0.24</td>
</tr>
<tr>
<td>Total</td>
<td>3000</td>
<td>1.00</td>
</tr>
</tbody>
</table>

[You can use calculator lists to quickly calculate relative frequencies for a frequency table. See Calculator Note 5A.]

Jack: Whew! Now I see you must be right. The relative frequencies from the simulation match your probabilities fairly well. I admit I fell down a bit here. I forgot that there are two ways to get one person choosing correctly—the first chooses correctly and the second doesn't, or the second chooses correctly and the first doesn't. There is only one way for two people to choose correctly: The first person chooses correctly and the second person chooses correctly.

Jill: Now that this has been settled, let's try to construct the probability distribution for three people, or \( n = 3 \).

Jack: Let me redeem myself. Following your reasoning that different orders should be listed separately, there are eight possible outcomes.

<table>
<thead>
<tr>
<th>First Person</th>
<th>Second Person</th>
<th>Third Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T )</td>
<td>( T )</td>
<td>( T )</td>
</tr>
<tr>
<td>( T )</td>
<td>( T )</td>
<td>( B )</td>
</tr>
<tr>
<td>( T )</td>
<td>( B )</td>
<td>( T )</td>
</tr>
<tr>
<td>( B )</td>
<td>( T )</td>
<td>( T )</td>
</tr>
<tr>
<td>( T )</td>
<td>( B )</td>
<td>( B )</td>
</tr>
<tr>
<td>( B )</td>
<td>( T )</td>
<td>( B )</td>
</tr>
<tr>
<td>( B )</td>
<td>( B )</td>
<td>( T )</td>
</tr>
<tr>
<td>( B )</td>
<td>( B )</td>
<td>( B )</td>
</tr>
</tbody>
</table>
Jill: I’ll count the number of people in each of the eight equally likely outcomes who picked the tap water. That gives us the probability distribution in Display 5.3. The probability that all three people choose correctly is $\frac{1}{8}$.

<table>
<thead>
<tr>
<th>Number Who Choose T</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$\frac{1}{8}$</td>
</tr>
<tr>
<td>1</td>
<td>$\frac{3}{8}$</td>
</tr>
<tr>
<td>2</td>
<td>$\frac{3}{8}$</td>
</tr>
<tr>
<td>3</td>
<td>$\frac{1}{8}$</td>
</tr>
</tbody>
</table>

Display 5.3  Jack and Jill’s probability distribution for $n = 3$.

Jack: This makes me think we need a larger sample size. Even if nobody can identify tap water, there is a one-eighth chance everyone will choose correctly! The reputation of Downhill Research is at stake!

**Sample Spaces**

Jack and Jill both used the same principle: Start by making a list of possible outcomes. Over the years, mathematicians realized that such a list of possible outcomes, called a sample space, must satisfy specific requirements.

A **sample space** for a chance process is a complete list of disjoint outcomes. All of the outcomes in a sample space must have a total probability equal to 1.

*Complete* means that every possible outcome is on the list. *Disjoint* means that two different outcomes can’t occur on the same opportunity. Sometimes the term *mutually exclusive* is used instead of *disjoint*. This book will alternate between the two terms so you can get used to both of them.

Deciding whether two outcomes are disjoint sounds easy enough, but be careful. You have to think about what an outcome means in your situation. If your outcome is the result of a single coin flip, your sample space is heads ($H$) and tails ($T$). These two outcomes are mutually exclusive because you can’t get both heads and tails on a single flip. But suppose you are thinking about what happens when you flip a coin three times. Now your sample space includes outcomes like $HHT$ and $TTT$. Even though you get tails on the third flip in both $HHT$ and $TTT$, these are disjoint outcomes. Your sample space consists of triples of flips, and these aren’t the same triple.

Jack’s sample space—neither person chooses $T$, one chooses $T$, two choose $T$—has outcomes that are complete and disjoint, but they aren’t equally likely. He has a legitimate sample space; he has simply assigned the wrong probabilities.

**DISCUSSION**

**Sample Spaces**

D4. In a slightly different form, Jack and Jill’s argument holds an important place in the early history of probability. One famous French mathematician,
Jean d’Alembert (1717–83), wrote in a 1750s text that the probability of getting two heads in two flips of a fair coin is $\frac{1}{3}$. Here’s his list of outcomes, which he said were equally likely:

- The first flip is tails.
- The first flip is heads, the second is tails.
- The first flip is heads, the second is heads.

Is his list of outcomes complete? Are the outcomes disjoint? Are the outcomes equally likely?

Data and Symmetry

How can you tell if the outcomes in your sample space are equally likely? There’s no magic answer, only two useful principles—data and symmetry. A die has six sides, and the only thing that makes one side different from another is the number of dots. It’s hard to imagine that the number of dots would have much effect on which side lands facing up, so it’s reasonable to assume that the six sides are equally likely to land facing up. This seems like a good model for a die, but to verify it, you need data: Roll the die many times and compare your model’s predictions with the actual results to see if you have a good fit.

In Jack and Jill’s model, they are temporarily assuming that symmetry applies—that people are as likely to choose the tap water as the bottled water. In fact, it’s probably not the right model. When Jack and Jill collect data from enough real people, they might be able to reject it. Or perhaps the model is correct and they won’t be able to reject it. In either case, they will have done what they were hired to do—decide if people can distinguish the tap water from the bottled water.

Activity 5.1a will help you see some of the difficulties involved in coming up with a realistic probability model.

ACTIVITY 5.1a

Spinning Pennies

**What you’ll need:** one penny per student

In flipping a penny, heads and tails have the same probability. Is the same probability model true for spinning a penny?

1. Use one finger to hold your penny on edge on a flat surface, with Lincoln’s head right side up, facing you. Flick the penny with the index finger of your other hand as you let go so that it spins around many times on its edge. When it falls over, record whether it lands heads up or tails up.

2. Repeat, and combine results with the rest of your class until you have a total of at least 500 spins.

3. Are the data consistent with a model in which heads and tails are equally likely outcomes, or do you think the model can safely be rejected?

(continued)
4. Suppose you spin a penny three times and record whether it lands heads up or tails up.
   a. How many possible outcomes are there?
   b. Are these outcomes equally likely? If not, which is most likely? Least likely?

The Law of Large Numbers

Why was Jack convinced that his probability model was wrong after simulating the taste test with 3000 flips of two coins (Display 5.2 on page 290)? In Activity 5.1a, why might you be convinced that the equal probability model does not work for coin spinning? Results from observed data seem to preclude the proposed model. That is as it should be, according to a mathematical principle—the Law of Large Numbers—established in the 1600s by Swiss mathematician Jacob Bernoulli (1654–1705).

The Law of Large Numbers says that in random sampling, the larger the sample, the closer the proportion of successes in the sample tends to be to the proportion in the population. In other words, the difference between a sample proportion and the population proportion must get smaller (except in rare instances) as the sample size gets larger. You can see a demonstration of this law in Display 5.4. Think of spinning a coin in such a way that it comes up heads with probability 0.4. The graph in Display 5.4 shows a total of 150 spins, with the proportion of heads accumulated after each spin plotted against the number of the spin. Notice that the coin came up heads for the first few spins, so the proportion of heads starts out at 1. This proportion quickly decreases, however, as the number of spins increases, and it ends up at about 0.377 after 150 spins.

Display 5.4  Proportion of heads for the given number of spins of a coin, with \( P(\text{heads}) = 0.4 \).

[You can use a calculator to perform this experiment yourself. See Calculator Note 5B.]

Most people intuitively understand the Law of Large Numbers. If they want to estimate a proportion, they know it is better to take a larger sample than a smaller one. After Jack saw the results from 3000 flips of two coins, he immediately rejected his model that there are three equally likely outcomes. If there had been only 10 pairs of coin flips, he couldn't have been so sure that his model was wrong.
The Law of Large Numbers

D5. An opinion pollster says, “All I need to do to ensure the accuracy of the results of my polls is to make sure I have a large sample.”

A casino operator says, “All I need to do to ensure that the house will win most of the time is to keep a large number of people flocking into my casino.”

A manufacturer says, “All I need to do to keep my proportion of defective light bulbs low is to manufacture a lot of light bulbs.”

Comment on the correctness of each of these statements. In particular, do the people speaking understand the Law of Large Numbers?

D6. Flip a coin 20 times, keeping track of the cumulative number of heads and the cumulative proportion of heads after each flip. Plot the cumulative number of heads versus the flip number on one graph. Plot the cumulative proportion of heads versus the flip number on another graph. On each graph, connect the points. Repeat this process five times so that you have five sets of connected points on each plot. Comment on the differences in the patterns formed by the five lines on each graph.

The Fundamental Principle of Counting

Jack and Jill are now ready to tackle samples of size 5. They are beginning to realize that using Jill’s method will result in a long list of possible outcomes, and they have become a bit discouraged. The key is for them to notice these two things:

1. They can always think of asking their question in stages (one person at a time).
2. They can list the possible outcomes at each stage.

These two observations will let them count outcomes by multiplying. To see how, note that Jack and Jill also could have used a tree diagram like the one in Display 5.5 to count the number of outcomes for $n = 2$.

```
Person 1   Person 2   Outcome
T          T          • TT
T          B          • TB
B          T          • BT
B          B          • BB
```

Display 5.5 A tree diagram of all possible outcomes for $n = 2$. 
There are two main branches, one for each of the two ways the first person can answer. Each main branch has two secondary branches, one for each way the second person can answer after the first person has answered. In all, there are $2 \cdot 2$, or 4, possible outcomes.

For $n = 3$, you would add two branches to the end of each of the four branches in Display 5.5. This would give a total of $2 \cdot 2 \cdot 2$, or 8, possible outcomes, as Jack and Jill already discovered. This idea is called the **Fundamental Principle of Counting**.

**Fundamental Principle of Counting**

For a two-stage process with $n_1$ possible outcomes for stage 1 and $n_2$ possible outcomes for stage 2, the number of possible outcomes for the two stages taken together is $n_1n_2$.

More generally, if there are $k$ stages, with $n_i$ possible outcomes for stage $i$, then the number of possible outcomes for all $k$ stages taken together is $n_1n_2n_3 \cdots n_k$.

With five people, or $n = 5$, and two possible outcomes for each person, Jack and Jill have $2 \cdot 2 \cdot 2 \cdot 2 \cdot 2$, or 32, possible outcomes. If their model is correct—that is, if people are just guessing which is the tap water—these 32 outcomes are all equally likely. So the probability that each of the five people will correctly choose the tap water is $\frac{1}{32}$.

When a process has only two stages, it is often more convenient to list them using a two-way table.

**Example: Rolling Two Dice**

Make a two-way table that shows all possible outcomes when you roll two fair dice. What is the probability that you get doubles (both dice show the same number)?

**Solution**

Because there are six faces on each die and the dice are fair, there are $6 \cdot 6$, or 36, equally likely outcomes, as shown in Display 5.6. Six of these outcomes are doubles, so the probability of rolling doubles is $\frac{6}{36}$.

<table>
<thead>
<tr>
<th>First Roll</th>
<th>Second Roll</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4 5 6</td>
<td>1, 1 1, 2 1, 3 1, 4 1, 5 1, 6</td>
</tr>
<tr>
<td>2 1 2 2 2 3 2 4 2 5 2 6</td>
<td>2, 1 2, 2 2, 3 2, 4 2, 5 2, 6</td>
</tr>
<tr>
<td>3 1 3 2 3 3 3 4 3 5 3 6</td>
<td>3, 1 3, 2 3, 3 3, 4 3, 5 3, 6</td>
</tr>
<tr>
<td>4 1 4 2 4 3 4 4 4 5 4 6</td>
<td>4, 1 4, 2 4, 3 4, 4 4, 5 4, 6</td>
</tr>
<tr>
<td>5 1 5 2 5 3 5 4 5 5 5 6</td>
<td>5, 1 5, 2 5, 3 5, 4 5, 5 5, 6</td>
</tr>
<tr>
<td>6 1 6 2 6 3 6 4 6 5 6 6</td>
<td>6, 1 6, 2 6, 3 6, 4 6, 5 6, 6</td>
</tr>
</tbody>
</table>

Display 5.6 The 36 outcomes when rolling two dice.
The Fundamental Principle of Counting

D7. Suppose Jack and Jill have five people taste three well-known brands of cola. The people are asked if they prefer the first cola, the second cola, or the third cola or can't tell any difference. How many possible outcomes will be in Jack and Jill's list for this taste test? Are these outcomes equally likely?

D8. Suppose you flip a fair coin seven times.
   a. How many possible outcomes are there?
   b. What is the probability that you will get seven heads?
   c. What is the probability that you will get heads six times and tails once?

Summary 5.1: Constructing Models of Random Behavior

A probability model is a sample space together with an assignment of probabilities. The sample space is a complete list of disjoint outcomes for which these properties hold:

- Each outcome is assigned a probability between 0 and 1.
- The sum of all the probabilities is 1.

Often you can rely on symmetry to recognize that outcomes are equally likely. If they are, then you can compute probabilities simply by counting outcomes: The probability of an event is the number of outcomes that make up the event divided by the total number of possible outcomes. In statistics, the main practical applications of equally likely outcomes are in the study of random samples and in randomized experiments. As you saw in Chapter 4, in a survey, all possible simple random samples are equally likely. Similarly, in a completely randomized experiment, all possible assignments of treatments to units are equally likely.

The only way to decide whether a probability model is a reasonable fit to a real situation is to compare probabilities derived from the model with probabilities estimated from observed data.

The Fundamental Principle of Counting says that if you have a process consisting of $k$ stages with $n_i$ outcomes for stage $i$, the number of outcomes for all $k$ stages taken together is $n_1n_2n_3\cdots n_k$.

Practice

Where Do Probabilities Come From?

P1. Suppose Jack and Jill use a sample of four people who can't tell the difference between tap water and bottled water.
   a. Construct the probability distribution for the number of people in the sample who would choose the tap water just by chance.
   b. What is the probability that all four people will identify the tap water correctly?
   c. Is four people a large enough sample to ease Jack's concern about the reputation of Downhill Research?
P2. Display 5.7 gives the actual low temperature (to the nearest 5°F) in Oklahoma City on days when the National Weather Service forecast was for a low temperature of 30°F.

<table>
<thead>
<tr>
<th>Actual Low Temperature (°F)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>25</td>
<td>8</td>
</tr>
<tr>
<td>30</td>
<td>13</td>
</tr>
<tr>
<td>35</td>
<td>3</td>
</tr>
<tr>
<td>40</td>
<td>1</td>
</tr>
</tbody>
</table>

P5. Jean d’Alembert (whom you met in D4) was coauthor (with another Frenchman, Denis Diderot, 1713–84) of a 35-volume Encyclopédie. In it, he wrote that the probability of getting heads at least once in two flips of a fair coin is \( \frac{2}{3} \). This time, he said that these three outcomes were equally likely:
- heads on the first flip
- heads on the second flip
- heads on neither flip

a. Is this list of outcomes complete?
b. Are the outcomes disjoint?
c. Are the three outcomes equally likely?
d. Is d’Alembert correct about the probability of getting heads at least once?
The Law of Large Numbers

P7. The results of 50 spins of a penny are plotted in Display 5.8. The horizontal axis gives the number of the spin, and the vertical axis gives the cumulative proportion of spins so far that were heads.

a. Was the first spin heads or tails? The second? The third? The fourth? The 50th?

b. Use these data to estimate the probability that this penny will land heads up when it is spun.

Display 5.8 Results of 50 spins of a penny.

The Fundamental Principle of Counting

P8. Suppose you ask a person to taste a particular brand of strawberry ice cream and evaluate it as good, okay, or poor on flavor and as acceptable or unacceptable on price.

a. Show all possible outcomes on a tree diagram.

b. How many possible outcomes are there?

c. Are all the outcomes equally likely?

P9. A dental clinic has three dentists and seven dental hygienists.

a. If you are assigned a dentist and a dental hygienist at random when you go in, how many different pairs could you end up with?

b. What is the probability that you get your favorite dentist and your favorite dental hygienist?

c. Illustrate your answer in part a with a two-way table.

d. Illustrate your answer in part a with a tree diagram.

Exercises

E1. Suppose you flip a coin five times and count the number of heads.

a. List all possible outcomes.

b. Make a table that gives the probability distribution for the number of heads.

c. What is the probability that you get at most four heads?

E2. Refer to the sample space for rolling two dice shown in Display 5.6 on page 295. Make a table that gives the probability distribution for the sum of the two dice. The first column should list the possible sums, and the second column should list their probabilities.

E3. Refer to the sample space for rolling two dice shown in Display 5.6 on page 295. Determine each of these probabilities.

a. not getting doubles

b. getting a sum of 5

c. getting a sum of 7 or 11

d. a 5 occurring on the first die

e. getting at least one 5

f. a 5 occurring on both dice

g. the larger number is a 5 (if you roll doubles, the number is both the smaller number and the larger number)

h. the smaller number is a 5

i. the difference of the larger number and the smaller number is 5
E4. A tetrahedral die has four sides, with the numbers 1, 2, 3, and 4.

a. How many equally likely outcomes are there when two fair tetrahedral dice are rolled?
b. Using a two-way table, show all possible equally likely outcomes.
c. What is the probability of getting doubles?
d. What is the probability that the sum is 2?
e. What is the probability that the larger number is a 2? (If you roll doubles, the number is both the smaller number and the larger number.)
f. What is the probability that the smaller number is a 2?
g. What is the probability that the larger number minus the smaller number is equal to 2?
h. What is the probability that the smaller number is less than 2?

E5. Suppose you roll a tetrahedral die twice (see E4). Determine whether each proposed sample space is disjoint and complete. If the sample space is disjoint and complete, assign probabilities to the given outcomes. If it is not, explain why not.

a. {no 4 on the two rolls, one 4, two 4’s}
b. {the first roll is a 4, the second roll is a 4, neither roll is a 4}
c. {the first 4 comes on the first roll, the first 4 comes on the second roll}
d. {the sum of the two rolls is less than 2, the sum is more than 2}
e. {the first roll is a 4, neither roll is a 4}

E6. Data or symmetry?

a. A taste test will be conducted to see if people prefer the taste of food A or food B. In fact, food A is identical in taste to food B. Is there any way to tell from symmetry alone (as opposed to needing data) that food A will be preferred over food B half the time? Give an argument in support of your answer.
b. Is there any way to tell from symmetry (as opposed to data) that a randomly selected mathematics major will be female with probability about 0.5? Give an argument in support of your answer.

E7. The results of 50 rolls of a pair of dice are plotted in Display 5.9. The horizontal axis gives the number of the roll, and the vertical axis gives the cumulative proportion of rolls so far that were doubles.

a. Was the first roll a double? The second? The third? The fourth?
b. How many rolls were doubles?
c. Use this sample to estimate the probability of rolling doubles with a pair of dice.
d. Why do the lengths of the line segments tend to get shorter as the number of spins increases?

E8. Business is slow at Downhill Research. Jack and Jill decide to investigate the probability that a penny they found will land heads up when spun.
a. Jack spins the penny 500 times and gets 227 heads. What is his estimate of the
probability that the penny will land heads up when spun?

b. Jill spins the penny 50 times. Would you expect Jack or Jill to have an estimate closer to the true probability of the penny landing heads up? Why?

c. Jack and Jill now flip the penny 10,000 times. Their results are recorded in Display 5.10. Fill in the missing percentages. Do their results illustrate the Law of Large Numbers? Explain.

<table>
<thead>
<tr>
<th>Number of Flips</th>
<th>Number of Heads</th>
<th>Percentage of Heads</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>6</td>
<td>60</td>
</tr>
<tr>
<td>100</td>
<td>56</td>
<td>56</td>
</tr>
<tr>
<td>1,000</td>
<td>524</td>
<td>—?—</td>
</tr>
<tr>
<td>10,000</td>
<td>5,140</td>
<td>—?—</td>
</tr>
</tbody>
</table>

Display 5.10 Percentage of heads for various numbers of coin flips.

d. Does the number of heads get closer to or farther away from half the number of flips?

E9. Miguel is designing an experiment to test which combination of colors makes the type on computer screens easiest to read. His two factors are the color of the background (blue, green, yellow, white, or beige) and the color of the type (brown, black, navy, or gray).

a. How many different possible treatments are there?

b. Show the possible treatments on a tree diagram.

c. Suppose Miguel adds a third factor, brightness, to his experiment. He will have two levels of brightness on the screen, high and low. How many possible distinct treatments are there?

d. Make a tree diagram showing this new situation.

E10. Tran has six shirts (blue, green, red, yellow, and two white—one long-sleeved, one short-sleeved) and four pairs of pants (brown, black, blue, and gray).

a. Use the Fundamental Principle of Counting to find the number of possible outfits Tran can wear.

b. Show all possible outfits in a two-way table.

c. If Tran selects a shirt and a pair of pants at random, determine the probability of Tran's wearing
   i. a white shirt
   ii. the gray pants
   iii. a white shirt and the gray pants
   iv. a white shirt or the gray pants

E11. In a Culver City, California election, the two candidates for a seat on the Board of Education tied with 1141 votes each. To avoid a long and costly run-off election, the city code called for a chance process to select the winner. The city officials made up two bags. One candidate was given a bag that contained eight red marbles and one white marble. The other candidate was given a bag that contained eight blue marbles and one white marble. Simultaneously, the candidates drew a marble. If neither marble was white, the candidates continued to draw marbles simultaneously, one marble at a time. The first candidate to draw a white marble was the winner. The winner drew a white marble on the fourth round. [Source: Los Angeles Times, December 2, 2003, pages B1, B7.]

a. Is this a fair process?

b. How could this process once again result in a tie? What's the chance of a tie on the first round?

c. How could you change the process to reduce the probability of a tie?

E12. You have 27 songs on your MP3 player and randomly select one to play. It's your favorite. When it finishes, you again randomly select a song to play, and it's the same one!

a. What is the chance this happening?

b. What is the chance that both random selections will be the same song?

E13. Equally likely outcomes?

a. Use the Fundamental Principle of Counting to find the number of outcomes for the situation of flipping a fair coin six times. Can you find the probability that you will get heads all six times?
b. Use the Fundamental Principle of Counting to find the number of outcomes for the situation of rolling a fair die six times. Can you find the probability that you will get a 3 all six times?

c. Use the Fundamental Principle of Counting to find the number of outcomes for the situation of picking 1200 U.S. residents at random and asking if they go to school. Can you find the probability that all 1200 will say yes?

E14. How many three-digit numbers can you make from the digits 1, 2, and 7? You can use the same digit more than once. If you choose the digits at random, what is the probability that the number is less than 250?

5.2 Using Simulation to Estimate Probabilities

In Activity 1.1a, you used simulation to estimate a probability. You wrote the ages of ten hourly workers on slips of paper, drew three slips to represent those workers to be laid off, and computed their average age. After repeating this process many times, you were able to estimate the probability of getting, just by chance, an average age as large as or even larger than the actual average age. In this section, you will learn a more efficient method of conducting a simulation, using random digits.

A table of random digits is a string of digits that is constructed in such a way that each digit has probability \( \frac{1}{10} \) of being 0, 1, 2, 3, 4, 5, 6, 7, 8, or 9. Further, each digit is selected independently of the previous digits. A table of random digits appears in Table D on page 828 of this book. When you need a string of random digits, you can use this table or you can generate a string of digits using your calculator. [See Calculator Note 4A.]

These steps, and the five examples that follow, illustrate how to use a table of random digits to estimate answers to questions about probability.

The Steps in a Simulation That Uses Random Digits

1. Assumptions. State the assumptions you are making about how the real-life situation works. Include any doubts you might have about the validity of your assumptions.

2. Model. Describe how you will use random digits to conduct one run of a simulation of the situation.
   - Make a table that shows how you will assign a digit (or a group of digits) to represent each possible outcome. (You can disregard some digits.)
   - Explain how you will use the digits to model the real-life situation. Tell what constitutes a single run and what summary statistic you will record.

3. Repetition. Run the simulation a large number of times, recording the results in a frequency table. You can stop when the distribution doesn't change to any significant degree when new results are included.

4. Conclusion. Write a conclusion in the context of the situation. Be sure to say that you have an estimated probability.
Example: Hourly Workers at Westvaco

The ages of the ten hourly workers involved in Round 2 of the layoffs were 25, 33, 35, 38, 48, 55, 55, 55, 56, and 64. The ages of the three workers who were laid off were 55, 55, and 64, with average age 58. Use simulation with random digits to estimate the probability that three workers selected at random for layoff would have an average age of 58 or more.

Solution

1. Assumptions.

You are assuming that each of the ten workers has the same chance of being laid off and that the workers to be laid off are selected at random without replacement.

2. Model.

- Assign each worker a random digit as in Display 5.11.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Digit Assigned</th>
</tr>
</thead>
<tbody>
<tr>
<td>The worker age 25</td>
<td>1</td>
</tr>
<tr>
<td>The worker age 33</td>
<td>2</td>
</tr>
<tr>
<td>The worker age 35</td>
<td>3</td>
</tr>
<tr>
<td>The worker age 38</td>
<td>4</td>
</tr>
<tr>
<td>The worker age 48</td>
<td>5</td>
</tr>
<tr>
<td>The first worker age 55</td>
<td>6</td>
</tr>
<tr>
<td>The second worker age 55</td>
<td>7</td>
</tr>
<tr>
<td>The third worker age 55</td>
<td>8</td>
</tr>
<tr>
<td>The worker age 56</td>
<td>9</td>
</tr>
<tr>
<td>The worker age 64</td>
<td>0</td>
</tr>
</tbody>
</table>

Display 5.11 Assignment of random numbers to Westvaco workers.

- Start at a random place in a table of random digits and look at the next three digits. Those digits will represent the workers selected to be laid off in a single run of the simulation. Because the same person can’t be laid off twice, if a digit repeats, ignore it and go to the next digit. Find the average of the ages of the three workers laid off.

3. Repetition.

Suppose the string of random digits begins like this:

32416 15000 56054

In the first run of the simulation, the digits 3, 2, and 4 represent the workers ages 35, 33, and 38. Record their average age, 35.33.

In the second run of the simulation, the digits 1, 6, and 5 represent the workers laid off. The second 1 was skipped because that worker was already selected in this run. The ages are 25, 55, and 48. Record the average age, 42.67.
In the third run of the simulation, the digits 0, 5, and 6 represent the workers laid off. The second and third 0’s were skipped because that worker was already selected in this run. The ages are 64, 48, and 55. Record the average age, 55.67. Display 5.12 shows the results of 2000 runs of this simulation, including the three described.

4. **Conclusion.**

From the histogram, the estimated probability of getting an average age of 58 years or more if you pick three workers at random is \( \frac{90}{2000} \) or 0.045. This probability is fairly small, so it is unlikely that the process Westvaco used for layoffs in this round was equivalent to picking the three workers at random.

According to the Law of Large Numbers, the more runs you do, the closer you can expect your estimated probability to be to the theoretical probability. The simulation in the previous example contained 2000 runs. When you do a simulation on a computer, there is no reason not to do many thousands of runs.

**Example: Hand Washing in Public Restrooms**

The American Society of Microbiology periodically estimates the percentage of people who wash their hands after using a public restroom. (Yes, these scientists spy on people in public restrooms, but it’s for a good cause.) They conducted the first study in 1996, when 67% of people washed their hands. [Source: Sports Illustrated, “Less Chop, More Soap,” September 22, 2005, www.SI.com.] Suppose you watched four randomly selected people using a public restroom back in 1996. Use simulation to estimate the probability that all four washed their hands.

**Solution**

1. **Assumptions.**

   You are assuming that the four people were selected at random and independently from the population of restroom users. That is, you didn’t sample by doing something like taking the next four people in one randomly selected public restroom.

   You also are assuming that 67% is the true percentage of all restroom users who wash their hands. This assumption may or may not be true, because the percentage was estimated from a (large) sample.
Summarizing, the situation to be modeled is taking a random sample of size 4 from a large population with 67% “hand-washers” and 33% “non-hand-washers.”

2. Model.

- Because there are two decimal places in 0.67, group the random digits into pairs. The pairs will be assigned to outcomes as in Display 5.13.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Pairs of Digits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Get a hand-washer</td>
<td>1–67</td>
</tr>
<tr>
<td>Get a non-hand-washer</td>
<td>68–99, 0</td>
</tr>
</tbody>
</table>

**Display 5.13** Assignment of random numbers to hand-washers and non-hand-washers.

- Look at four pairs of random digits. Count and record how many of the four pairs of digits represent hand-washers. This time, if a pair of digits repeats, use it again because each pair of digits doesn’t represent a specific person as in the previous example.

3. Repetition.

Suppose the string of random digits begins like this:

```
01420 94975 89283 40133 48486
```

In the first run of the simulation, the pairs 01, 42, 09, and 49 represent hand-washer, hand-washer, hand-washer, and hand-washer. Record a 4, for four hand-washers in this run.

In the second run of the simulation, the pairs 75, 89, 28, and 34 represent non-hand-washer, non-hand-washer, hand-washer, and hand-washer. Record a 2, for the two hand-washers in this run.

In the third run of the simulation, the pairs 01, 33, 48, and 48 represent four hand-washers. Record a 4, for the four hand-washers in this run. Note that the 48 was used twice in the same run because it doesn’t represent only a single hand-washer.

The frequency table and histogram in Display 5.14 give the results of 10,000 runs of this simulation, including the three already described.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 hand-washers</td>
<td>124</td>
</tr>
<tr>
<td>1 hand-washer</td>
<td>975</td>
</tr>
<tr>
<td>2 hand-washers</td>
<td>2,967</td>
</tr>
<tr>
<td>3 hand-washers</td>
<td>3,964</td>
</tr>
<tr>
<td>4 hand-washers</td>
<td>1,970</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>10,000</strong></td>
</tr>
</tbody>
</table>

**Display 5.14** Results of 10,000 runs of the hand-washer simulation.
4. Conclusion.
   From the frequency table, the estimated probability that all four randomly
   selected people will wash their hands is \( \frac{1,970}{10,000} \) or 0.197.

Example: Men and Women and Hand Washing

Things have improved. The American Society of Microbiology now reports that
75% of men and 90% of women wash their hands after using a public restroom.
[Source: www.harrisinteractive.com.] Suppose you pick a man and a woman at random as
they use a public restroom. Estimate the probability that both wash their hands.

Solution

1. Assumptions.
   You are assuming that the man and the woman were selected randomly and
   independently from the population of restroom users.
   Your second assumption is that 75% and 90% are the true percentages.
   That assumption may or may not be true, because those percentages were
   estimated from a (large) sample.

2. Model.
   In this example, you will learn how a calculator can be used to generate
   pairs of random integers to simulate this situation. Because there are two
decimal places in 0.75 and 0.90, you'll generate two integers from 0 to 99.
The first integer will represent a man, and the second will represent a
woman. The integers will be assigned to outcomes as in Display 5.15.

   \[
   \begin{array}{|l|l|}
   \hline
   \text{Outcome for Man} & \text{First Integer} \\
   \hline
   \text{Get a hand-washer} & 1–75 \\
   \text{Get a non-hand-washer} & 76–99, 0 \\
   \hline
   \text{Outcome for Woman} & \text{Second Integer} \\
   \hline
   \text{Get a hand-washer} & 1–90 \\
   \text{Get a non-hand-washer} & 91–99, 0 \\
   \hline
   \end{array}
   \]

   Display 5.15 Assignment of random integers to male and female
   hand-washers and non-hand-washers.

   • Record whether the first integer represents a man who washed or didn’t
     wash and whether the second integer represents a woman who washed or
     didn’t wash. If an integer repeats, use it again.

3. Repetition.
   A calculator gives this sequence of pairs of random integers. [See Calculator
   Note 4A to learn how to generate pairs of random integers.]
In the first run of the simulation, the integers 72 and 43 represent a man who washed and a woman who washed.

In the second run of the simulation, the integers 6 and 81 represent a man who washed and a woman who washed.

In the third run of the simulation, the integers 76 and 81 represent a man who didn’t wash and a woman who washed.

Display 5.16 gives the results of 5000 runs of this simulation, including the three already described. [See Calculator Note 5C to learn how to use your calculator to do many runs of a simulation and report the results.]

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Both washed</td>
<td>3361</td>
</tr>
<tr>
<td>Man washed/woman didn’t</td>
<td>360</td>
</tr>
<tr>
<td>Man didn’t wash/woman did</td>
<td>1155</td>
</tr>
<tr>
<td>Neither washed</td>
<td>124</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>5000</strong></td>
</tr>
</tbody>
</table>

Display 5.16 Results of 5000 runs of the male/female hand-washer simulation.

4. **Conclusion.**

From the frequency table, the estimated probability that both the randomly selected man and the randomly selected woman wash their hands is \( \frac{3361}{5000} \), or about 0.67.

**Example: Waiting for a Teen Who Flunks a Treadmill Test**


Suppose you want to interview a teen who flunks a treadmill test about his or her health habits. On average, to how many teens would you have to give a treadmill test before you find one who flunks?

**Solution**

1. **Assumptions.**

You are assuming that you are testing randomly and independently selected teens. You also are assuming that one-third is the true proportion of teens who would flunk a treadmill test.

2. **Model.**

- Assign the random digits as shown in Display 5.17. Ignore the digits 4 through 9 and 0.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Digits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teen flunks test</td>
<td>1</td>
</tr>
<tr>
<td>Teen passes test</td>
<td>2 and 3</td>
</tr>
</tbody>
</table>

Display 5.17 Assignment of random numbers to teens who flunk or pass a treadmill test.
• Look at random digits until you find a 1, which represents a teen who flunks the treadmill test. Record how many digits you have to look at in order to get a 1. Don’t count any digits 4 through 9 or 0.

3. **Repetition.**

Suppose the string of random digits begins like this:

\[ 22053 \quad 71491 \quad 32923 \quad 71926 \]

In the first run of the simulation, the digit 2 represents a teen who passes the test. The second digit also represents a teen who passes. Skip the 0 and the 5. The 3 also represents a teen who passes. Skip the 7. The 1 represents a teen who flunks. Record a 4, because you had to test four teens to find one who flunked.

In the second run of the simulation, skip the 4 and the 9. The 1 represents a teen who flunks the test. Record a 1, because you had to test only one teen to find one who flunked.

In the third run of the simulation, the 3 and the 2 represent teens who pass. Skip the 9. The 2 and the 3 represent teens who pass. Skip the 7. The 1 represents a teen who flunks. Record a 5, because you had to test five teens to find one who flunked.

The histogram and frequency table in Display 5.18 give the results of 3000 runs of this simulation, including the three described.

To compute the average number of teens who have to be tested, use the formula from page 67. The average is about 3.

4. **Conclusion.**

The estimate from this simulation is that, on average, you will have to test about three teens to find one who flunks the treadmill test.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 tested</td>
<td>967</td>
</tr>
<tr>
<td>2 tested</td>
<td>651</td>
</tr>
<tr>
<td>3 tested</td>
<td>434</td>
</tr>
<tr>
<td>4 tested</td>
<td>299</td>
</tr>
<tr>
<td>5 tested</td>
<td>221</td>
</tr>
<tr>
<td>6 tested</td>
<td>135</td>
</tr>
<tr>
<td>7 tested</td>
<td>95</td>
</tr>
<tr>
<td>8 tested</td>
<td>71</td>
</tr>
<tr>
<td>9 tested</td>
<td>51</td>
</tr>
<tr>
<td>10 tested</td>
<td>25</td>
</tr>
<tr>
<td>11 tested</td>
<td>12</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>3000</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 tested</td>
<td>9</td>
</tr>
<tr>
<td>13 tested</td>
<td>13</td>
</tr>
<tr>
<td>14 tested</td>
<td>6</td>
</tr>
<tr>
<td>15 tested</td>
<td>4</td>
</tr>
<tr>
<td>16 tested</td>
<td>3</td>
</tr>
<tr>
<td>17 tested</td>
<td>1</td>
</tr>
<tr>
<td>18 tested</td>
<td>1</td>
</tr>
<tr>
<td>19 tested</td>
<td>0</td>
</tr>
<tr>
<td>20 tested</td>
<td>0</td>
</tr>
<tr>
<td>21 tested</td>
<td>2</td>
</tr>
<tr>
<td>22 tested</td>
<td>0</td>
</tr>
</tbody>
</table>

Display 5.18  Results of 3000 runs of the treadmill simulation.
Example: Collecting Blends of Coffee

A coffee house rotates daily among six different blends of coffee. Suppose Sydney goes into the coffee house on three randomly selected days. Estimate the probability that the coffee house will be offering a different blend on each of the three days.

Solution

1. Assumptions.
   You are assuming that each time Sydney goes into the coffee house, the probability is \( \frac{1}{6} \) that any given blend will be offered and that her three trips are selected independently of one another.

2. Model.
   • Assign the digits 1 through 6 to the six blends of coffee, as in Display 5.19.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Digit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blend 1</td>
<td>1</td>
</tr>
<tr>
<td>Blend 2</td>
<td>2</td>
</tr>
<tr>
<td>Blend 3</td>
<td>3</td>
</tr>
<tr>
<td>Blend 4</td>
<td>4</td>
</tr>
<tr>
<td>Blend 5</td>
<td>5</td>
</tr>
<tr>
<td>Blend 6</td>
<td>6</td>
</tr>
</tbody>
</table>

   Display 5.19 Assignment of random numbers to blends of coffee.

   • This time you will practice using a calculator to generate three random integers between 1 and 6. Record whether the three digits are different.

3. Repetition.
   A calculator gives these sets of random integers. [See Calculator Note 4A.]

   The first run of the simulation resulted in 1, 1, and 2. Because Sydney got Blend 1 twice, record that the three blends weren’t all different.
   In the second run of the simulation, the three numbers are different, so record that Sydney got three different blends.
   In the third run of the simulation, the three numbers are different, so again record that Sydney got three different blends.
   Display 5.20 gives the results of 2000 runs of this simulation.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>The three blends weren’t all different.</td>
<td>884</td>
</tr>
<tr>
<td>The three blends were all different.</td>
<td>1116</td>
</tr>
<tr>
<td>Total</td>
<td>2000</td>
</tr>
</tbody>
</table>

   Display 5.20 Results of 2000 runs of the coffee blend simulation.
4. **Conclusion.**

From the frequency table, the estimated probability that Sydney will get three different blends is \( \frac{1116}{2000} \) or about 0.56.

---

**Using Simulation to Estimate Probabilities**

D9. What characteristics would you expect a table of random digits to have?

D10. Jason needs a table of random digits with the digits selected from 1 to 6. He has a fair six-sided die that has the digits 1 through 6 on its faces. To construct his table, he rolls the die and writes down the number that appears on the top of the die for the first random digit. To speed things up, Jason writes the number that appears on the bottom of the die for the next random digit. For example, if a 2 appears on the top of the die, a 5 would be on the bottom, so Jason’s sequence of digits would begin 2 5. He continues in this way.

a. Does each of the digits 1 through 6 have an equally likely chance of appearing in any given position in Jason’s table?

b. Are the digits selected independently of previous digits?

c. Has Jason constructed a table of random digits?

D11. For the example Hand Washing in Public Restrooms, on page 303, describe another way the random digits could have been assigned using pairs of random digits. Describe a way that uses triples of random digits.

D12. Describe how to use a die to simulate the situation in the coffee house example.

**Summary 5.2: Using Simulation to Estimate Probabilities**

In this section, you have learned to model various situations involving probability and to estimate their solutions using simulation. Some of the problems you worked on were amusing, but don’t let this mislead you. Simulation is an important method of estimating answers to problems that are too complicated to solve theoretically.

You have used random digits to model situations because it is quick and easy to generate random digits on calculators and computers. However, many situations may be easily modeled using other chance devices such as coins, dice, or spinners. The key to designing a simulation is to be clear about what an outcome on the chance device corresponds to in the real-life situation.

When you solve a problem using simulation, always include these four steps:

1. State your assumptions.
2. Describe how you will use random digits to conduct one run of the simulation.
3. Run the simulation a large number of times, recording the results in a frequency table.
4. Write a conclusion in the context of the situation.
Using Simulation to Estimate Probabilities

P10. How would you use a table of random digits to conduct one run of a simulation of each situation?

a. There are eight workers, ages 27, 29, 31, 34, 35, 42, and 47. Three are to be chosen at random for layoff.

b. There are 11 workers, ages 27, 29, 31, 34, 35, 42, 42, 42, 46, and 47. Four are to be chosen at random for layoff.

For P11–P13, complete a–d.


b. Model. Make a table that shows how you are assigning the random digits to the outcomes. Explain how you will use the digits to model the situation and what summary statistic you will record.

c. Repetition. Conduct ten runs of the simulation, using the specified row of Table D on page 828. Add your results to the frequency table given in the practice problem.

d. Conclusion. Write a conclusion in the context of the situation.

P11. Researchers at the Macfarlane Burnet Institute for Medical Research and Public Health in Melbourne, Australia, noticed that the teaspoons had disappeared from their tearoom. They purchased new teaspoons, numbered them, and found that 80% disappeared within 5 months.


Suppose that 80% is the correct probability that a teaspoon will disappear within 5 months and that this group purchases ten new teaspoons. Estimate the probability that all the new teaspoons will be gone in 5 months.

Start at the beginning of row 34 of Table D on page 828, and add your ten results to the frequency table in Display 5.21, which gives the results of 4990 runs.

Display 5.21 Results of 4990 runs of the disappearing teaspoons.

P12. A catastrophic accident is one that involves severe skull or spinal damage. The National Center for Catastrophic Sports Injury Research reports that over the last 21 years, there have been 101 catastrophic accidents among female high school and college athletes. Fifty-five of these resulted from cheerleading.

[Source: www.unc.edu.]

Suppose you want to study catastrophic accidents in more detail, and you take a random sample, without replacement, of 8 of these 101 accidents. Estimate the probability that at least half of your eight sampled accidents resulted from cheerleading.

Start at the beginning of row 17 of Table D on page 828, and add your ten runs to the frequency table in Display 5.22, which gives the results of 990 runs.
P13. The winner of the World Series of baseball is the first team to win four games. That means the series can be over in four games or can go as many as seven games. Suppose the two teams playing are equally matched. Estimate the probability that the World Series will go seven games before there is a winner.

Start at the beginning of row 9 of Table D on page 828, and add your ten runs to the frequency table in Display 5.23, which gives the results of 4990 runs.

<table>
<thead>
<tr>
<th>Number of Games in World Series</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>610</td>
</tr>
<tr>
<td>5</td>
<td>1303</td>
</tr>
<tr>
<td>6</td>
<td>1541</td>
</tr>
<tr>
<td>7</td>
<td>1536</td>
</tr>
<tr>
<td>Total</td>
<td>4990</td>
</tr>
</tbody>
</table>

Display 5.23 Results of 4990 runs of the number of games in the World Series.

Exercises

For E15–E20, complete a–d.


b. Model. Make a table that shows how you are assigning the random digits to the outcomes. Explain how you will use the digits to conduct one run of the simulation and what summary statistic you will record.

c. Repetition. Conduct ten runs of the simulation, using the specified row of Table D on page 828. Add your results to the frequency table given in the exercise.

d. Conclusion. Write a conclusion in the context of the situation.

E15. About 10% of high school girls report that they rarely or never wear a seat belt while riding in motor vehicles. Suppose you randomly sample four high school girls. Estimate the probability that no more than one of the girls says that she rarely or never wears a seat belt. [Source: www.nhtsa.dot.gov]

Start at the beginning of row 36 of Table D on page 828, and add your ten results to the frequency table in Display 5.24, which gives the results of 9990 runs.

<table>
<thead>
<tr>
<th>Number of Girls Who Rarely or Never Wear a Seat Belt</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6635</td>
</tr>
<tr>
<td>1</td>
<td>2861</td>
</tr>
<tr>
<td>2</td>
<td>462</td>
</tr>
<tr>
<td>3</td>
<td>32</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>9990</td>
</tr>
</tbody>
</table>

Display 5.24 Results of 9990 runs of the number of girls, out of four, who rarely or never wear a seat belt.
E16. The study cited in E15 says that 18% of high school boys report that they rarely or never wear a seat belt. Suppose you randomly sample nine high school boys. Estimate the probability that none of the boys say that they rarely or never wear a seat belt.

Start at the beginning of row 41 of Table D on page 828, and add your ten results to the frequency table in Display 5.25, which gives the results of 9990 runs.

<table>
<thead>
<tr>
<th>Number of Boys Who Rarely or Never Wear a Seat Belt</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1646</td>
</tr>
<tr>
<td>1</td>
<td>3223</td>
</tr>
<tr>
<td>2</td>
<td>2966</td>
</tr>
<tr>
<td>3</td>
<td>1533</td>
</tr>
<tr>
<td>4</td>
<td>505</td>
</tr>
<tr>
<td>5</td>
<td>113</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>9990</td>
</tr>
</tbody>
</table>

Display 5.25 Results of 9990 runs of the number of boys, out of nine, who rarely or never wear a seat belt.

E17. You forgot to study for a ten-question multiple-choice test on the habits of the three-toed sloth. You will have to guess on each question. Each question has four possible answers. Estimate the probability that you will answer half or more of the questions correctly.

Start at the beginning of row 35 of Table D on page 828, and add your ten results to the frequency table in Display 5.26, which gives the results of 5990 runs.

<table>
<thead>
<tr>
<th>Number of Correct Answers</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>356</td>
</tr>
<tr>
<td>1</td>
<td>1162</td>
</tr>
<tr>
<td>2</td>
<td>1700</td>
</tr>
<tr>
<td>3</td>
<td>1464</td>
</tr>
<tr>
<td>4</td>
<td>840</td>
</tr>
<tr>
<td>5</td>
<td>352</td>
</tr>
<tr>
<td>6</td>
<td>89</td>
</tr>
<tr>
<td>7</td>
<td>24</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>5990</td>
</tr>
</tbody>
</table>

Display 5.26 Results of 5990 runs of the number of correct answers, out of ten guesses.

E18. A Harris poll estimated that 25% of U.S. residents believe in astrology. Suppose you would like to interview a person who believes in astrology. Estimate the probability that you will have to ask four or more U.S. residents to find one who believes in astrology. [Source: www.sciencedaily.com.]

Start at the beginning of row 15 of Table D on page 828, and add your ten results to the frequency table in Display 5.27, which gives the results of 1990 runs.

Display 5.27 Results of 1990 runs of the number of U.S. residents who believe in astrology.

312 Chapter 5 Probability Models
<table>
<thead>
<tr>
<th>Number of People Asked</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>437</td>
</tr>
<tr>
<td>2</td>
<td>412</td>
</tr>
<tr>
<td>3</td>
<td>278</td>
</tr>
<tr>
<td>4</td>
<td>210</td>
</tr>
<tr>
<td>5</td>
<td>173</td>
</tr>
<tr>
<td>6</td>
<td>131</td>
</tr>
<tr>
<td>7</td>
<td>94</td>
</tr>
<tr>
<td>8</td>
<td>70</td>
</tr>
<tr>
<td>9</td>
<td>52</td>
</tr>
<tr>
<td>10</td>
<td>37</td>
</tr>
<tr>
<td>11</td>
<td>24</td>
</tr>
<tr>
<td>12</td>
<td>15</td>
</tr>
<tr>
<td>13</td>
<td>14</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of Babies</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>941</td>
</tr>
<tr>
<td>2</td>
<td>480</td>
</tr>
<tr>
<td>3</td>
<td>265</td>
</tr>
<tr>
<td>4</td>
<td>156</td>
</tr>
<tr>
<td>5</td>
<td>60</td>
</tr>
<tr>
<td>6</td>
<td>51</td>
</tr>
<tr>
<td>7</td>
<td>17</td>
</tr>
<tr>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>1990</td>
</tr>
</tbody>
</table>

**Display 5.27** Results of 1990 runs of the number of people asked in order to find the first who believes in astrology.

E19. The probability that a baby is a girl is about 0.49. Suppose a large number of couples each plan to have babies until they have a girl. Estimate the average number of babies per couple.

Start at the beginning of row 9 of Table D on page 828, and add your ten results to the frequency table in Display 5.28, which gives the results of 1990 runs.

<table>
<thead>
<tr>
<th>Number of Babies</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>886</td>
</tr>
<tr>
<td>4</td>
<td>1394</td>
</tr>
<tr>
<td>5</td>
<td>1485</td>
</tr>
<tr>
<td>6</td>
<td>1347</td>
</tr>
<tr>
<td>7</td>
<td>1136</td>
</tr>
<tr>
<td>8</td>
<td>903</td>
</tr>
<tr>
<td>9</td>
<td>676</td>
</tr>
<tr>
<td>10</td>
<td>491</td>
</tr>
<tr>
<td>11</td>
<td>368</td>
</tr>
<tr>
<td>12</td>
<td>330</td>
</tr>
<tr>
<td>13</td>
<td>263</td>
</tr>
<tr>
<td>14</td>
<td>167</td>
</tr>
<tr>
<td>15</td>
<td>137</td>
</tr>
<tr>
<td>16</td>
<td>103</td>
</tr>
<tr>
<td>17</td>
<td>81</td>
</tr>
<tr>
<td>18</td>
<td>67</td>
</tr>
<tr>
<td>19</td>
<td>44</td>
</tr>
<tr>
<td>20</td>
<td>29</td>
</tr>
<tr>
<td>21</td>
<td>20</td>
</tr>
<tr>
<td>22</td>
<td>14</td>
</tr>
<tr>
<td>23</td>
<td>7</td>
</tr>
<tr>
<td>24</td>
<td>12</td>
</tr>
<tr>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>26</td>
<td>5</td>
</tr>
<tr>
<td>27 or more</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>9990</td>
</tr>
</tbody>
</table>

**Display 5.28** Results of 1990 runs of the number of babies born to a couple who have babies until they have a girl.

E20. Boxes of cereal often have small prizes in them. Suppose each box of one type of cereal contains one of four different small cars. Estimate the average number of boxes a parent will have to buy until his or her child gets all four cars.

Start at the beginning of row 49 of Table D on page 828, and add your ten results to the frequency table in Display 5.29, which gives the results of 9990 runs.

<table>
<thead>
<tr>
<th>Number of Boxes</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>886</td>
</tr>
<tr>
<td>5</td>
<td>1394</td>
</tr>
<tr>
<td>6</td>
<td>1485</td>
</tr>
<tr>
<td>7</td>
<td>1347</td>
</tr>
<tr>
<td>8</td>
<td>1136</td>
</tr>
<tr>
<td>9</td>
<td>903</td>
</tr>
<tr>
<td>10</td>
<td>676</td>
</tr>
<tr>
<td>11</td>
<td>491</td>
</tr>
<tr>
<td>12</td>
<td>368</td>
</tr>
<tr>
<td>13</td>
<td>330</td>
</tr>
<tr>
<td>14</td>
<td>263</td>
</tr>
<tr>
<td>15</td>
<td>167</td>
</tr>
<tr>
<td>16</td>
<td>137</td>
</tr>
<tr>
<td>17</td>
<td>103</td>
</tr>
<tr>
<td>18</td>
<td>81</td>
</tr>
<tr>
<td>19</td>
<td>67</td>
</tr>
<tr>
<td>20</td>
<td>44</td>
</tr>
<tr>
<td>21</td>
<td>29</td>
</tr>
<tr>
<td>22</td>
<td>20</td>
</tr>
<tr>
<td>23</td>
<td>14</td>
</tr>
<tr>
<td>24</td>
<td>7</td>
</tr>
<tr>
<td>25</td>
<td>12</td>
</tr>
<tr>
<td>26</td>
<td>5</td>
</tr>
<tr>
<td>27 or more</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>9990</td>
</tr>
</tbody>
</table>

**Display 5.29** Results of 9990 runs of the number of cereal boxes bought by parents who buy boxes until their child gets all four cars.
For E21–E26, complete a–d.

a. **Assumptions.** State your assumptions.

b. **Model.** Make a table that shows how you are assigning the random digits to the outcomes. Explain how you will use the digits to model the situation and what summary statistic you will record.

c. **Repetition.** Conduct 20 runs of the simulation, using the specified row of Table D on page 828. Make your own frequency table.

d. **Conclusion.** Write a conclusion in the context of the situation.

E21. A group of five friends have backpacks that all look alike. They toss their backpacks on the ground and later pick up a backpack at random. Estimate the probability that everyone gets his or her own backpack. Start at the beginning of row 31 of Table D on page 828.

E22. One method of testing whether a person has extrasensory perception (ESP) involves a set of 25 cards, called Zener cards. Five cards have a circle printed on them, five a square, five a star, five a plus sign, and five have wavy lines (see Display 5.30). The experimenter shuffles the cards and concentrates on one at a time. The subject is supposed to identify which card the experimenter is looking at. Once the experimenter looks at a card, it is not reused (and the subject is not told what it is). Estimate the probability that a subject will identify the first five cards correctly just by guessing circle, square, star, plus sign, and wavy lines, in that order. Start at the beginning of row 41 of Table D on page 828.

E23. There are six different car keys in a drawer, including yours. Suppose you grab one key at a time until you get your car key. Estimate the probability that you get your car key on the second try. Start at the beginning of row 13 of Table D on page 828.

E24. A deck of 52 cards contains 13 hearts. Suppose you draw cards one at a time, without replacement. Estimate the probability that it takes you four cards or more to draw the first heart. Start at the beginning of row 28 of Table D on page 828.

E25. Every so often, a question about probability gets the whole country stirred up. These problems involve a situation that is easy to understand, but they have a solution that seems unreasonable to most people. The most famous of these problems is the so-called “Monty Hall problem” or “the problem of the car and the goats.” To put it mildly, tempers have flared over this problem. Here’s the situation: At the end of the television show *Let’s Make a Deal*, a contestant would be offered a choice of three doors. Behind one door was a good prize, say a car. Behind the other two doors were lesser prizes, say goats. After the contestant chose a door, the host, Monty Hall, would open one of the two doors that the contestant hadn’t chosen. The opened door always had a goat. Once, after the door with a goat had been opened, the contestant asked if he now could switch from his (unopened) door to the other unopened door. Was this a good strategy for the contestant, was it a bad strategy, or did it make no difference? Start at the beginning of row 8 of Table D on page 828.

Display 5.30 Zener cards.
E26. This question appeared in the “Ask Marilyn” column.
I work at a waste-treatment plant, and we do assessments of the time-to-failure and time-to-repair of the equipment, then input those figures into a computer model to make plans. But when I need to explain the process to people in other departments, I find it difficult. Say a component has two failure modes. One occurs every 5 years, and the other occurs every 10 years. People usually say that the time-to-failure is 7.5 years, but this is incorrect. It’s between 3 and 4 years. Do you know of a way to explain this that people will accept? [Source: PARADE, March 13, 2005, p. 17.]

Use simulation to estimate the expected time-to-failure for this component. Start at the beginning of row 40 of Table D on page 828.

5.3 The Addition Rule and Disjoint Events

One of the shortest words in the English language—*or*—is often misunderstood. This is because *or* can have two different meanings. For example, if you are told at a party that you can have apple pie or chocolate cake, you might not be sure whether you must pick only one or whether it would be acceptable to say “Both!”

In statistics, the meaning of *or* allows *both* as a possibility: You can have apple pie or you can have chocolate cake or you can have both. Thus, in statistics, to find \( P(A \text{ or } B) \), you must find the probability that \( A \) happens or \( B \) happens or both \( A \) and \( B \) happen.

### Disjoint and Complete Categories

Statistical data often are presented in tables like the one in Display 5.31, which gives basic figures for employment in the United States for all nonmilitary adults who were employed or seeking employment.

<table>
<thead>
<tr>
<th>Nonmilitary Employable Adults</th>
<th>Number of People (in thousands)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employees on farm payrolls</td>
<td>8,975</td>
</tr>
<tr>
<td>Employees on nonfarm payrolls</td>
<td>135,354</td>
</tr>
<tr>
<td>Seeking employment</td>
<td>7,205</td>
</tr>
<tr>
<td>Total civilian labor force</td>
<td>151,534</td>
</tr>
</tbody>
</table>

**Display 5.31** The U.S. labor force as of July 2006.

The categories are disjoint, which makes computing probabilities easy. For example, if you pick one of the people in the civilian labor force at random, you can find the probability that he or she is employed by adding the employees on farm payrolls to those on nonfarm payrolls and then dividing by the total number in the civilian labor force:

\[
P(\text{employed}) = P(\text{on farm payroll or on nonfarm payroll})
\]

\[
= \frac{8,975 + 135,354}{151,534} = \frac{144,329}{151,534} \approx 0.952
\]
However, in the next example, you will see that you can’t always compute the probability of $A$ or $B$ simply by adding.

**Example: Leisure Activities of American Adults**

Display 5.32 lists some common leisure activities and the percentage of the 213 million U.S. adults who participated in the activity at least once in the prior 12 months.

a. Are these categories disjoint for categorizing an adult’s leisure activities? Are they complete?

b. How many adults read books for a leisure activity?

c. Can you find the percentage of U.S. adults who surfed the net or went to the beach?

<table>
<thead>
<tr>
<th>Activity</th>
<th>Percentage of U.S. Adults Who Engaged in Activity at Least Once in the Prior 12 Months</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dining out</td>
<td>48</td>
</tr>
<tr>
<td>Reading books</td>
<td>36</td>
</tr>
<tr>
<td>Surfing the net</td>
<td>27</td>
</tr>
<tr>
<td>Going to the beach</td>
<td>22</td>
</tr>
</tbody>
</table>

**Display 5.32** Percentage of U.S. adults who engaged in various leisure activities in 2004.  
(Source: U.S. Census Bureau, *Statistical Abstract of the United States*, 2006, Table 1227.)

**Solution**

a. These categories aren’t disjoint because, for example, the same person might have dined out and read books. Also, for this table, you can tell that the categories aren’t disjoint because the percentages sum to more than 100%.

b. Thirty-six percent of 213,000,000, or 76,680,000 adults, read books for leisure.

c. From this information, it is impossible to determine the percentage of U.S. adults who surfed the net or went to the beach. If you add the two percentages, you are counting the people twice who did both. Because you don’t know how many people that is, you are stuck.

**The Addition Rule for Disjoint Events**

While doing Activity 5.3a, notice when you can find $P(A \text{ or } B)$ simply by adding.

---

**ACTIVITY 5.3a**

**Exploring or**

1. For each question, record the number of students in your class who answer yes.
   - Is blue your favorite color?
   - Is red your favorite color?

(continued)
2. From the information in step 1 alone, can you determine the number of students in your class who say blue is their favorite color or who say red is their favorite color? If so, explain how. If not, what additional information do you need?

3. For each of these questions, record the number of students in your class who answer yes.
   - Are you female?
   - Is blue your favorite color?

4. From the information in step 3 alone, can you determine the number of students in your class who are female or who have blue as their favorite color? If so, explain how. If not, what additional information do you need?

In Activity 5.3a, you can get the answer to one question by adding because the categories are disjoint. For the other question, adding doesn’t work because the categories are not disjoint.

Two useful rules emerge from Activity 5.3a. First, if two events are disjoint, the probability of their occurring together is 0.

**A Property of Disjoint (Mutually Exclusive) Events**

If event A and event B are disjoint, then

\[ P(A \text{ and } B) = 0 \]

Second, if two events are disjoint, then you can add probabilities to find 

\[ P(A \text{ or } B) \]

**The Addition Rule for Disjoint (Mutually Exclusive) Events**

If event A and event B are disjoint, then

\[ P(A \text{ or } B) = P(A) + P(B) \]

The Addition Rule for Disjoint Events can be generalized. For example, if each pair of events A, B, and C is disjoint, then

\[ P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C) \]

**Example: The Addition Rule with Dice**

Use the Addition Rule for Disjoint Events to find the probability that if you roll two dice, you get a sum of 7 or a sum of 8.
Solution
Let \( A \) be the event sum of 7. Let \( B \) be the event sum of 8. On the same roll, these two events are mutually exclusive, so \( P(A \text{ and } B) = 0 \). Then \( P(A \text{ or } B) = P(A) + P(B) \). Using the sample space of equally likely outcomes from Display 5.6 on page 295,

\[
P(\text{sum of 7 or sum of 8}) = P(\text{sum of 7}) + P(\text{sum of 8})
\]

\[
= \frac{6}{36} + \frac{5}{36} = \frac{11}{36}
\]

Example: The Addition Rule with Tap Water
Use the Addition Rule for Disjoint Events to compute the probability that exactly one person out of two correctly chooses the tap water in Jack and Jill’s model.

Solution
\[
P(\text{exactly one } T \text{ among two tasters})
\]

\[
= P(\{1st \text{ chooses } T \text{ and } 2nd \text{ chooses } B\} \text{ or } \{1st \text{ chooses } B \text{ and } 2nd \text{ chooses } T\})
\]

\[
= P(TB \text{ or } BT) = P(TB) + P(BT) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}
\]

You can add the probabilities because the two outcomes \( TB \) and \( BT \) are disjoint.

DISCUSSION

The Addition Rule for Disjoint Events
D13. Suppose you select a person at random from your school. Are these pairs of events mutually exclusive?

a. has ridden a roller coaster; has ridden a Ferris wheel
b. owns a classical music CD; owns a jazz CD
c. is a senior; is a junior
d. has brown hair; has brown eyes
e. is left-handed; is right-handed
f. has shoulder-length hair; is male

D14. If you select a person at random from your classroom, what is the probability that the person is a junior or a senior?

D15. Suppose there is a 20% chance of getting a mosquito bite each time you go outside on a summer evening. Can you use the Addition Rule for Disjoint Events to compute the probability that you will get bitten if you go outside three times? If you go outside six times?
The Addition Rule

What about computing $P(A \text{ or } B)$ when you don’t have disjoint categories? In that situation, addressed in Activity 5.3b, you must add an extra step.

**ACTIVITY 5.3b**

**More on or**

1. For each question, record the number of students in your class who answer yes.
   - Are you female?
   - Is blue your favorite color?
   - Is it true that you are female and blue is your favorite color?

2. From the information in step 1 alone, can you determine the number of students in your class who are female or have blue as their favorite color? (Remember, when we use the word *or* in probability, we mean *one or the other or both.*) If so, explain how to do it. If not, what additional information do you need?

3. Copy and complete this two-way table showing the number of members of your class whose responses fall into each cell of the table.

<table>
<thead>
<tr>
<th>Favorite Color</th>
<th>Blue</th>
<th>Not Blue</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4. Find the number of students who are female or have blue as their favorite color
   a. as a row total plus a column total minus a cell count
   b. as the sum of three cell counts

5. Let $B$ be the event that the favorite color of the randomly chosen student is blue. Let $F$ be the event that the randomly chosen student is female. Write two formulas for computing $P(B \text{ or } F)$, one corresponding to part a and one to part b in step 4. Use each formula to find $P(B \text{ or } F)$.

The ideas you explored in Activity 5.3b can be stated formally as the Addition Rule.

**Addition Rule**

For any two events $A$ and $B$,

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Using set notation, this rule is written

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

where $A \cap B$ is read "$A$ intersect $B".$$
Example: The Addition Rule for Events That Aren't Disjoint

Use the Addition Rule to find the probability that if you roll two dice, you get doubles or a sum of 8.

Solution

Let \( A \) be the event getting doubles, and let \( B \) be the event getting a sum of 8. These two events are not mutually exclusive, because they have a common outcome \((4, 4)\), which occurs with probability \( \frac{1}{36} \). From the Addition Rule, 
\[
P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B).
\]

For \( P(doubles \text{ or } sum \text{ of } 8) \):
\[
= P(doubles) + P(sum \text{ of } 8) - P(doubles \text{ and } sum \text{ of } 8)
\]
\[
= \frac{1}{36} + \frac{5}{36} - \frac{1}{36} = \frac{10}{36}
\]

You can also apply the Addition Rule “backward”; that is, you can compute \( P(A \text{ and } B) \) when you know the other probabilities.

Example: Computing \( P(A \text{ and } B) \)

In a local school, 80% of the students carry a backpack \((B)\) or a wallet \((W)\). Forty percent carry a backpack, and 50% carry a wallet. If a student is selected at random, find the probability that the student carries both a backpack and a wallet.

Solution

\( B \) and \( W \) aren’t disjoint, so the general form of the Addition Rule applies:
\[
P(B \text{ or } W) = P(B) + P(W) - P(B \text{ and } W)
\]
\[
0.80 = 0.40 + 0.50 - P(B \text{ and } W)
\]

Solving, \( P(B \text{ and } W) = 0.10 \). The probability that the student carries both a backpack and a wallet is 0.10.

The Addition Rule

D16. These diagrams are called Venn diagrams:

A. [Diagram A]
   B. [Diagram B]

   a. Which diagram illustrates mutually exclusive events?
   b. Use these diagrams to justify the two forms of the Addition Rule.
D17. What happens if you use the general form of the Addition Rule,

\[ P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \]

in a situation where \( A \) and \( B \) are mutually exclusive?

**Summary 5.3: The Addition Rule and Disjoint Events**

In statistics, \( P(A \text{ or } B) \) indicates the probability that event \( A \) occurs, event \( B \) occurs, or both occur in the same random outcome. The word or always allows for the possibility that both events occur. In this section, you learned the Addition Rule for any events \( A \) and \( B \):

\[ P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \]

Two events are disjoint (mutually exclusive) if they cannot occur on the same opportunity. In the case of disjoint events, \( P(A \text{ and } B) = 0 \), so the Addition Rule simplifies to

\[ P(A \text{ or } B) = P(A) + P(B) \]

Two-way tables often are helpful in clarifying the structure of a probability problem and in answering questions about the joint behavior of two events.

**Practice**

**Disjoint and Complete Categories**

P14. Of the 34,071,000 people in the United States who fish, 1,847,000 fish in the Great Lakes, 27,913,000 fish in other fresh water, and 9,051,000 fish in salt water. [Source: U.S. Census Bureau, Statistical Abstract of the United States, 2006, Table 1241.]

a. In categorizing people who fish, are these three categories disjoint? Are they complete?

b. Suppose you randomly select a person from among those who fish. Can you find the probability that the person fishes in salt water?

c. Suppose you randomly select a person from among those who fish. Can you find the probability that the person fishes in fresh water?

d. The number of people who fish in fresh water is 28,439,000. How many people fish in both salt water and fresh water?

P15. Display 5.33 categorizes the child support received by custodial parents with children under age 21 in the United States.

<table>
<thead>
<tr>
<th>Child Support Status by Custodial Parents in 2001</th>
<th>Number (in thousands)</th>
</tr>
</thead>
<tbody>
<tr>
<td>With child support agreement or award</td>
<td>7,916</td>
</tr>
<tr>
<td>Supposed to receive payments</td>
<td>6,924</td>
</tr>
<tr>
<td>Actually received payments</td>
<td>5,119</td>
</tr>
<tr>
<td>Received full amount</td>
<td>3,099</td>
</tr>
<tr>
<td>Received partial payments</td>
<td>2,020</td>
</tr>
<tr>
<td>Did not receive payments</td>
<td>1,805</td>
</tr>
<tr>
<td>Child support not awarded</td>
<td>5,467</td>
</tr>
<tr>
<td>Total custodial parents with children under age 21</td>
<td>13,383</td>
</tr>
</tbody>
</table>

Display 5.33 Custodial parents and court-ordered child support, 2001. [Source: U.S. Census Bureau, Statistical Abstract of the United States, 2006, Table 558.]

a. Revise the table so that the categories are complete and disjoint. Note that one category wasn’t included: custodial parents with a child support agreement or award who were not supposed to receive payments in 2001 (but maybe get them in some other year).
b. If one of the 13,383,000 custodial parents is selected at random, what is the probability that he or she was supposed to receive payments and received the full amount or received a partial amount?

The Addition Rule for Disjoint Events

P16. If you roll two dice, are these pairs of events mutually exclusive? Explain.
   a. doubles; sum is 8
   b. doubles; sum is odd
   c. a 3 on one die; sum is 10
   d. a 3 on one die; doubles

P17. A researcher will select a student at random from a school population where 33% of the students are freshmen, 27% are sophomores, 25% are juniors, and 15% are seniors.
   a. Is it appropriate to use the Addition Rule for Disjoint Events to find the probability that the student will be a junior or a senior? Why or why not?
   b. Find the probability that the student will be a freshman or a sophomore.

P18. A tetrahedral die has the numbers 1, 2, 3, and 4 on its faces. Suppose you roll a pair of tetrahedral dice.
   a. Make a table of all 16 possible outcomes (or use the one you made in E4 on page 299).
   b. Use the Addition Rule for Disjoint Events to find the probability that you get a sum of 6 or a sum of 7.
   c. Use the Addition Rule for Disjoint Events to find the probability that you get doubles or a sum of 7.
   d. Why can’t you use the Addition Rule for Disjoint Events to find the probability that you get doubles or a sum of 6?

The Addition Rule

P19. Display 5.34 gives information about all reportable crashes on state-maintained roads in North Carolina in a recent year.

<table>
<thead>
<tr>
<th></th>
<th>No Teen Driver(s)</th>
<th>Teen Driver(s)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Speed Related</td>
<td>120,082</td>
<td>24,291</td>
<td>144,373</td>
</tr>
<tr>
<td>Speed Related</td>
<td>67,331</td>
<td>19,755</td>
<td>87,086</td>
</tr>
<tr>
<td>Total</td>
<td>187,413</td>
<td>44,046</td>
<td>231,459</td>
</tr>
</tbody>
</table>

Display 5.34 Information on reportable crashes in North Carolina. [Source: North Carolina Crash Data Query, www.hsrc.unc.edu.]

   a. Are the events crash involved a teen driver and crash was speed related mutually exclusive? How can you tell?
   b. Use numbers from the cells of the table to compute the probability that a randomly selected crash involved a teen driver or was speed related.
   c. Now use two of the marginal totals and one number from a cell of the table to compute the probability that a randomly selected crash involved a teen driver or was speed related.

P20. Use the Addition Rule to compute the probability that if you roll two six-sided dice, 
   a. you get doubles or a sum of 4
   b. you get doubles or a sum of 7
   c. you get a 5 on the first die or you get a 5 on the second die

P21. Use the Addition Rule to compute the probability that if you flip two fair coins, you get heads on the first coin or you get heads on the second coin.

P22. Use the Addition Rule to find the probability that if you roll a pair of dice, you do not get doubles or you get a sum of 8.
Exercises

E27. Suppose you flip a fair coin five times. Are these pairs of events disjoint?
   a. You get five heads; you get four heads and one tails.
   b. The first flip is heads; the second flip is heads.
   c. You get five heads; the second flip is tails.
   d. You get three heads and two tails; the second flip is tails.
   e. The first four flips are heads; the first three flips are heads.
   f. Heads first occurs on the third flip; the fourth flip is heads.
   g. Heads first occurs on the third flip; three of the flips are heads.

E28. Suppose you select two people at random from a nearby health clinic. Are these pairs of events disjoint?
   a. One person has health insurance; the other person has health insurance.
   b. Both people have health insurance; only one person has health insurance.
   c. One person is over age 65; the other person is under age 19.
   d. One person is over age 95; the other person is the first person's mother.

E29. Display 5.35 shows the U.S. population categorized by age and sex, as reported by the U.S. Bureau of the Census. You are working for a polling organization that is about to select a random sample of U.S. residents.
   a. What is the probability that the first person selected will be female? Female and age 85 or older? Female or age 85 or older?
   b. What is the probability that the first person selected will be male? Male and under age 30? Male or under age 30?
   c. What proportion of females are age 85 or older? What proportion of people age 85 or older are female?

<table>
<thead>
<tr>
<th>Age in Years</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>17 and under</td>
<td>37,237,364</td>
<td>35,686,940</td>
<td>72,924,304</td>
</tr>
<tr>
<td>18 to 29</td>
<td>22,695,612</td>
<td>22,308,750</td>
<td>45,004,362</td>
</tr>
<tr>
<td>30 to 64</td>
<td>65,303,026</td>
<td>68,254,508</td>
<td>133,557,534</td>
</tr>
<tr>
<td>65 to 84</td>
<td>13,340,864</td>
<td>17,238,577</td>
<td>30,579,441</td>
</tr>
<tr>
<td>85 and older</td>
<td>1,205,952</td>
<td>2,419,908</td>
<td>3,625,860</td>
</tr>
<tr>
<td>Total</td>
<td>139,782,818</td>
<td>145,908,683</td>
<td>285,691,501</td>
</tr>
</tbody>
</table>

Display 5.35 Resident population of the United States by age and sex (excluding the population living in institutions, college dormitories, and other group quarters), 2004. [Source: U.S. Census Bureau, 2004 American Community Survey, Table B01001.]

E30. Display 5.36 shows the U.S. college population categorized by age and sex.

<table>
<thead>
<tr>
<th>Age in Years</th>
<th>Male Enrollment (in thousands)</th>
<th>Female Enrollment (in thousands)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>19 and under</td>
<td>1,629</td>
<td>2,033</td>
<td>3,662</td>
</tr>
<tr>
<td>20 to 24</td>
<td>3,129</td>
<td>3,724</td>
<td>6,853</td>
</tr>
<tr>
<td>25 to 29</td>
<td>982</td>
<td>1,181</td>
<td>2,163</td>
</tr>
<tr>
<td>30 and older</td>
<td>1,577</td>
<td>2,383</td>
<td>3,960</td>
</tr>
<tr>
<td>Total</td>
<td>7,317</td>
<td>9,321</td>
<td>16,638</td>
</tr>
</tbody>
</table>

Display 5.36 College population of the United States by age and sex. [Source: U.S. Census Bureau, Statistical Abstract of the United States, 2006, Table 270.]

a. What percentage of college students are female? Female and age 30 or older? Female or age 30 or older?
b. What percentage of college students are male? Male and under age 30? Male or under age 30?
c. What proportion of female college students are age 30 or older? What proportion of college students age 30 or older are females?
E31. Display 5.37 describes students in a large class.

<table>
<thead>
<tr>
<th>Owns a Laptop?</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Has a Dog?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>46</td>
<td></td>
<td>72</td>
</tr>
<tr>
<td>No</td>
<td>112</td>
<td>72</td>
<td>216</td>
</tr>
</tbody>
</table>

Display 5.37 Table of laptop ownership and dog possession.

a. Fill in the missing cells and marginal totals.
b. What is the probability that a student randomly selected from this class doesn't have a dog or doesn't own a laptop?

E32. Display 5.38 classifies crashes in a recent year in Virginia. Some cells are missing because they weren't given directly by the source.

<table>
<thead>
<tr>
<th>Driver Violated a Traffic Law</th>
<th>Driver Didn’t Violate a Traffic Law</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fatality</td>
<td>521</td>
<td>837</td>
</tr>
<tr>
<td>No Fatality</td>
<td>—?—</td>
<td>—?—</td>
</tr>
<tr>
<td>Total</td>
<td>145,288</td>
<td>153,907</td>
</tr>
</tbody>
</table>

Display 5.38 Characteristics of crashes in Virginia. [Source: Virginia Department of Motor Vehicles, www.dmv.state.va.us.]

a. Fill in the missing cells and marginal totals.
b. What proportion of crashes involved a fatality and a traffic law violation?
c. What proportion of crashes involved a fatality or a traffic law violation?

d. What is the probability that the student carries both a backpack and a wallet. (Note that the word only makes this exercise different from the example on page 320.)

E35. In the United States, about 7,243,000 people ages 16 to 24 have completed high school and aren’t enrolled in college. There are 3,766,000 high school dropouts in this same age group. Of the high school graduates, 71% are employed, 11% are unemployed, and 18% are not in the labor force. Of the high school dropouts, 53% are employed, 14% are unemployed, and 33% are not in the labor force. If you randomly select a person age 16 to 24 who isn’t enrolled in college or high school, what is the probability that he or she is employed? Make a two-way table to answer this question. [Source: U.S. Census Bureau, Statistical Abstract of the United States, 2006, Table 261.]

E36. Polls of registered voters often report the percentage of Democrats and the percentage of Republicans who approve of the job the president is doing. Suppose that in a poll of 1500 randomly selected voters 860 are Democrats and 640 are Republicans. Overall, 937 approve of the job the president is doing, and 449 of these are Republicans. Assuming the people in this poll are representative, what percentage of registered voters are Republicans or approve of the job the president is doing? First answer this question by using the Addition Rule. Then make a two-way table showing the situation.

E37. Jill computes the probability that she gets heads at least once in two flips of a fair coin:

\[
P(\text{at least one heads in two flips}) = P(\text{heads on first flip or heads on second}) = P(\text{heads on first}) + P(\text{heads on second}) = \frac{1}{2} + \frac{1}{2} = 1
\]

She defends her use of the Addition Rule because getting heads on the first flip and heads on the second flip are mutually exclusive—both can’t happen at the same time. What would you say to her?
E38. Jill computes the probability that she gets heads exactly once in two flips of a fair coin:

\[ P(\text{exactly one heads in two flips}) \]
\[ = P(\text{tails on first flip and heads on second}) \]
\[ + P(\text{heads on first flip and tails on second}) \]
\[ = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \]

She defends her use of the Addition Rule because \( HT \) and \( TH \) are mutually exclusive. What would you say to her?

E39. When is this statement true?

\[ P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C) \]

E40. Suppose events \( A \), \( B \), and \( C \) are three events where \( P(A \text{ and } B) \neq 0 \), \( P(A \text{ and } C) \neq 0 \), \( P(B \text{ and } C) \neq 0 \), and \( P(A \text{ and } B \text{ and } C) \neq 0 \).

a. Draw a Venn diagram to illustrate this situation.

b. Use the Venn diagram from part a to help you write a rule for computing \( P(A \text{ or } B \text{ or } C) \).

---

### 5.4 Conditional Probability

The *Titanic* sank in 1912 without enough lifeboats for the passengers and crew. Almost 1500 people died, most of them men. Was that because a man was less likely than a woman to survive, or did more men die simply because men outnumbered women by more than 3 to 1 on the *Titanic*? You might turn to an old campfire song to decide what might have happened. One line is about loading the lifeboats: “And the captain shouted, ‘Women and children first.’” Statisticians aren’t opposed to songs. But statisticians also know that stories and theories, even stories set to music, are no substitute for data. Do the data in Display 5.39 support the lyrics of the song?

<table>
<thead>
<tr>
<th>Gender</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survived? Yes</td>
<td>367</td>
<td>344</td>
<td>711</td>
</tr>
<tr>
<td>No</td>
<td>1364</td>
<td>126</td>
<td>1490</td>
</tr>
<tr>
<td>Total</td>
<td>1731</td>
<td>470</td>
<td>2201</td>
</tr>
</tbody>
</table>

Display 5.39 *Titanic* survival data. [Source: Journal of Statistics Education 3, no. 3 (1995).]

Although the numbers alone can’t tell you who got to go first on the lifeboats, they do show that \( \frac{344}{470} \) or roughly 73%, of the females survived, while only \( \frac{367}{1731} \) or roughly 21%, of the males survived. Thus, the data are fully consistent with the hypothesis that the song explains what happened.

Overall, \( \frac{711}{2201} \) or approximately 32.3%, of the people survived, but the survival rate for females was much higher and that for males much lower. The chance of surviving depended on the *condition* of whether the person was male or female. This commonsense notion that probability can change if you are given additional information is called **conditional probability**.
Conditional Probability from the Sample Space

Often you can calculate conditional probabilities directly from the sample space.

**Example: The Titanic and Conditional Probability**

Suppose you pick a person at random from the list of people aboard the *Titanic*. Let $S$ be the event that this person survived, and let $F$ be the event that the person was female. Find $P(S|F)$, read “the probability of $S$ given that $F$ is known to have happened.”

**Solution**

The conditional probability $P(S|F)$ is the probability that the person survived *given the condition* that the person was female. To find $P(S|F)$, restrict your sample space to only the 470 females (the outcomes for whom the condition is true). Then compute the probability of survival as the number of favorable outcomes—344, the women surviving—divided by the total number of outcomes—470—in the restricted sample space.

$$P(S|F) = P(\text{survived} | \text{female}) = \frac{344}{470} \approx 0.732$$

Overall, only 32.3% of the people survived, but 73.2% of the females survived.

**Example: Sampling Without Replacement**

When you sample without replacement from a small population, the probabilities for the second draw depend on what happens on the first draw. For example, imagine a population of four recent graduates (that is, $N = 4$), two from the west ($W$) and two from the east ($E$). Suppose you randomly choose one person from the population and he or she is from the west. What is the probability that the second person selected is also from the west? In symbols, find

$$P(\text{W selected 2nd} | \text{W selected 1st})$$

**Solution**

If you start with a population {$W, W, E, E$} and the first selection is $W$, that leaves the restricted sample space {$W, E, E$} from which to choose on the second draw. The conditional probability that the second selection is from the west is

$$P(\text{W selected 2nd} | \text{W selected 1st}) = \frac{1}{3}$$

**DISCUSSION**

Conditional Probability from the Sample Space

D18. When you compare sampling with and without replacement, how does the size of the population affect the comparison? Conditional probability lets you answer the question quantitatively. Imagine two populations of students, one large ($N = 100$) and one small ($N = 4$), with half the students
in each population male ($M$). Draw random samples of size $n = 2$ from each population.

a. First, consider the small population. Find $P(2\text{nd } M|1\text{st } M)$, assuming you sample without replacement. Then calculate the probability again, this time with replacement.

b. Do the same for the larger population.

c. Based on these calculations, how would you describe the effect of population size on the difference between the two sampling methods?

The Multiplication Rule for $P(A \text{ and } B)$

You will now investigate a general rule for finding $P(A \text{ and } B)$ for any two events $A$ and $B$. Suppose, for example, you choose one person at random from the passenger list of the *Titanic*. What is the probability that the person was female ($F$) and survived ($S$)? You already know how to answer this question directly by using the data in Display 5.39. Out of 2201 passengers, 344 were women and survived, so

$$P(F \text{ and } S) = \frac{344}{2201}$$

You can also find the answer using conditional probability. It sometimes helps to show the order of events in a tree diagram, such as the one in Display 5.40.

[Tree diagram for the Titanic data]

You can interpret the top branch of the tree diagram, for example, as

$$P(F) \cdot P(S|F) = \frac{470}{2201} \cdot \frac{344}{470} = \frac{344}{2201} = P(F \text{ and } S)$$

The tree diagram leads to a rule for finding $P(A \text{ and } B)$, shown in the box on the next page.
The Multiplication Rule

The probability that event $A$ and event $B$ both happen is given by

$$P(A \text{ and } B) = P(A) \cdot P(B \mid A)$$

Alternatively,

$$P(A \text{ and } B) = P(B) \cdot P(A \mid B)$$

Display 5.41 shows the Multiplication Rule on the branches of a tree diagram.

From past work with fractions, you know that $of$ translates to multiplication. Three-fourths of 20 means $\frac{3}{4} \cdot 20$, or 15. The Multiplication Rule is simply an extension of this same logic.

<table>
<thead>
<tr>
<th>Event</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$ and $B$</td>
<td>$P(A) \cdot P(B \mid A)$</td>
</tr>
<tr>
<td>$B$ and $A$</td>
<td>$P(B) \cdot P(A \mid B)$</td>
</tr>
</tbody>
</table>

Display 5.41 The general Multiplication Rule, shown on a tree diagram.

Example: Females on the Titanic

Use the Multiplication Rule to compute the percentage of passengers on the Titanic who were females and survived.

Solution

Twenty-one percent of all Titanic passengers were female, and 73% of female passengers survived. So 73% of 21%, or 15.33%, of all passengers were female and survived.

$$P(\text{female and survived}) = (0.21)(0.73) = P(\text{female}) \cdot P(\text{survived|female})$$

Example: Coincidences

This article appeared in the *Los Angeles Times* on July 18, 1978.

**Man, Wife Beat Odds in Moose-Hunt Draw**

The Washington State Game Department conducted a public drawing last week in Olympia for three moose hunting permits. There were 2,898 application cards in the wire mesh barrel. It was cranked around
several times before the first name was drawn: Judy Schneider of East Wenatchee, Wash. The barrel was spun again. Second name drawn: Bill Schneider, Judy’s husband. Result: groans of disbelief from hopeful moose hunters in the auditorium.

a. What is the probability that Judy’s name would be the first drawn and her husband’s name would be second?

b. What assumptions are you making when you answer this question?

c. Comment on why coincidences like this seem to happen so often.

**Solution**

a. 
\[
P(\text{Judy first and Bill second}) = P(\text{Judy first}) \cdot P(\text{Bill second} | \text{Judy first})
\]
\[
= \frac{1}{2898} \cdot \frac{1}{2897} \approx 0.000000119
\]

b. The assumptions are that Bill and Judy had only one card each in the barrel and that the cards were well mixed.

c. The chance that Judy’s name and then her husband’s name will be called is about one chance in 10 million. However, there is also the chance that Bill’s name will be called first and then Judy’s. This doubles their chances of being the first two names drawn to 0.000000238—only about one chance in 5 million. However, suppose all the names in the barrel were those of couples. Then the probability that the first two names drawn will be those of a couple is \(\frac{1}{2897}\), or about 0.0000345, because the first name can be anyone and that person’s partner is one of the 2897 left for the second draw. This probability is still small, but now it is 3 chances in 10,000. Finally, thousands of lotteries take place in the United States every year, so it is virtually certain that coincidences like this will happen occasionally and be reported in the newspaper.

**DISCUSSION**

The Multiplication Rule for \(P(A \text{ and } B)\)

D19. Use the tree diagram for the Titanic data in Display 5.40 on page 327 to find \(P(\text{male and survived})\) and \(P(\text{male and died})\).

D20. Use the Titanic data in Display 5.39 on page 325 to show that

\[
P(M \text{ and } S) = P(S) \cdot P(M | S) = P(M) \cdot P(S | M)
\]

D21. Use the frequencies in the table below to show that

\[
P(A \text{ and } B) = P(A) \cdot P(B | A) = P(B) \cdot P(A | B)
\]

is true in general.

<table>
<thead>
<tr>
<th>Event A Present?</th>
<th>(c)</th>
<th>(d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event B Present?</td>
<td>(e)</td>
<td>(f)</td>
</tr>
</tbody>
</table>
The Definition of Conditional Probability

It's time to define conditional probability formally.

For any two events $A$ and $B$ such that $P(B) > 0$,

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

The Multiplication Rule is a consequence of the formal definition of conditional probability. Simply solve the equation in the definition for $P(A \text{ and } B)$.

**Example: Conditional Probability**

Suppose you roll two dice. Use the definition of conditional probability to find the probability that you get a sum of 8 given that you rolled doubles.

**Solution**

Thinking in terms of the restricted sample space, you already know that the answer is $\frac{1}{6}$ because there are six equally likely ways to roll doubles and only one of these is a sum of 8. Using the definition of conditional probability, the computation looks like this.

$$P(\text{sum 8} \mid \text{doubles}) = \frac{P(\text{sum 8 and doubles})}{P(\text{doubles})} = \frac{1}{36} = \frac{1}{6}$$

**Conditional Probability and Medical Tests**

In medicine, screening tests give a quick indication of whether a person is likely to have a particular disease. (For example, the ELISA test is a screening test for HIV, and a chest X ray is a screening test for lung cancer.) Because screening tests are intended to be relatively quick and noninvasive, they often are not as accurate as other tests that take longer or are more invasive. (For example, a biopsy is more accurate than a chest X ray if you want to know whether a lung tumor is cancerous.)

A two-way table like the one in Display 5.42 is often used to show the four possible outcomes of a screening test.

<table>
<thead>
<tr>
<th>Disease</th>
<th>Positive</th>
<th>Negative</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>a</td>
<td>b</td>
<td>a + b</td>
</tr>
<tr>
<td>Absent</td>
<td>c</td>
<td>d</td>
<td>c + d</td>
</tr>
<tr>
<td>Total</td>
<td>a + c</td>
<td>b + d</td>
<td>a + b + c + d</td>
</tr>
</tbody>
</table>

Display 5.42 Possible results of a screening test.
The effectiveness of screening tests is judged using conditional probabilities. These four terms are commonly used. In each case, higher values are better.

Positive predictive value (PPV) = \( P(\text{disease} | \text{test positive}) = \frac{a}{a + c} \)

Negative predictive value (NPV) = \( P(\text{no disease present} | \text{test negative}) = \frac{d}{b + d} \)

Sensitivity = \( P(\text{test positive} | \text{disease present}) = \frac{a}{a + b} \)

Specificity = \( P(\text{test negative} | \text{no disease}) = \frac{d}{c + d} \)

**Example: A Relatively Rare Disease**

Relatively rare diseases have tables similar to Display 5.43. This test has 50 false positives; that is, 50 people who don’t have the disease tested positive. It has only 1 false negative—a person with the disease who tested negative. Find the sensitivity, specificity, PPV, and NPV for this test.

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Positive</th>
<th>Negative</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Present</td>
<td>9</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Absent</td>
<td>50</td>
<td>9,940</td>
<td>9,990</td>
</tr>
<tr>
<td>Total</td>
<td>59</td>
<td>9,941</td>
<td>10,000</td>
</tr>
</tbody>
</table>

Display 5.43  Hypothetical results of a screening test for a rare disease.

**Solution**

In some ways, this is a pretty good test. It finds 9 out of the 10 people who have the disease, for a sensitivity of 0.9. It correctly categorizes 9940 out of the 9990 people who don’t have the disease, for a specificity of 0.99. The NPV is 9940 out of 9941, or 0.9999. However, notice that only 9 out of 59, or 15%, of the people who test positive for the disease actually have it! The PPV is quite low because there are so many false positives.

The previous example shows what can happen when the population being screened is mostly disease free, even with a test of high specificity. Most of the people who test positive do not, in fact, have the disease. On the other hand, if the population being screened has a high incidence of the disease, then there tend to be many false negatives and the negative predictive value tends to be low.

Because the positive predictive value and the negative predictive value depend on the population as well as the screening test, statisticians prefer to judge a test based on the other pair of conditional probabilities: sensitivity and specificity.

**DISCUSSION**

**Conditional Probability and Medical Tests**

D22. Most patients who take a screening test find the positive predictive value and the negative predictive value easier to interpret than sensitivity and specificity. Explain why.
Chapter 5: Probability Models

Conditional Probability and Statistical Inference

To calculate a probability, you must work from a model. For example, suppose you want to calculate the probability of observing an even number of dots on a roll of a die. Your model is that the die is fair. Using this model, you know that

$$P(\text{even number} | \text{fair die}) = \frac{3}{6}$$

Now suppose you have a die that you suspect isn't fair. If you roll it and get an even number, you cannot calculate $P(\text{fair die} | \text{even number})$. So how can you discredit a model?

If you roll the die 20 times and you get an even number every time, you can compute (by the method of the next section)

$$P(\text{even number on all 20 rolls} \mid \text{fair die}) = \left(\frac{3}{6}\right)^{20} \approx 0.000001$$

This outcome is so unlikely that you can feel justified in abandoning the model that the die is fair. But you still didn't—and can't—compute

$$P(\text{fair die} \mid \text{even number on all 20 rolls})$$

The dialogue that follows is invented and did not actually occur in the Westvaco case (Chapter 1), but it is based on real conversations one of your authors had on several occasions with a number of different lawyers as they grappled with conditional probabilities.

**Statistician:** Suppose you draw three workers at random from the set of ten hourly workers. This establishes random sampling as the model for the study.

**Lawyer:** Okay.

**Statistician:** It turns out that there are $\binom{10}{3}$, or 120, possible samples of size 3, and only 6 of them give an average age of 58 or more.

**Lawyer:** So the probability is $\frac{6}{120}$, or 0.05.

**Statistician:** Right.

**Lawyer:** There's only a 5% chance that the company didn't discriminate and a 95% chance that it did.

**Statistician:** No, that's not true.

**Lawyer:** But you said . . .

**Statistician:** I said that if the age-neutral model of random draws is correct, then there's only a 5% chance of getting an average age of 58 or more.

**Lawyer:** So the chance that the company is guilty must be 95%.
Statistician: Slow down. If you start by assuming the model is true, you can compute the chances of various results. But you're trying to start from the results and compute the chance that the model is right or wrong. You can't do that.

Here is the analysis: The lawyer is having trouble with conditional probabilities. The statistician has computed

\[ P(\text{average age} \geq 58 \mid \text{random draws}) = 0.05 \]

The lawyer wants to know

\[ P(\text{no discrimination} \mid \text{average age} = 58) \]

The statistician has computed \( P(\text{data} \mid \text{model}) \). The lawyer wants to know \( P(\text{model} \mid \text{data}) \). Finding the probability that there was no discrimination given that the average age was 58 is not a problem that statistics can solve. A model is needed in order to compute a probability.

**DISCUSSION**

Conditional Probability and Statistical Inference

D23. There's a parallel between statistical testing and medical testing. Write expressions for conditional probabilities—in terms of finding a company guilty or not guilty of age discrimination—that correspond to a PPV and an NPV. (You might think of the true medical status as comparable to the true state of affairs within the company.)

**Summary 5.4: Conditional Probability**

In this section, you saw these important uses of conditional probability in statistics:

- to compare sampling with and without replacement
- to study the effectiveness of medical screening tests
- to describe probabilities of the sort used in statistical tests of hypotheses

You have learned the definition of conditional probability and the Multiplication Rule and how to use these two key concepts in solving practical problems. When you are solving conditional probability problems, by using the Multiplication Rule or otherwise, tree diagrams and two-way tables help you organize the data and see the structure of the problem.

- The definition of conditional probability is that for any two events \( A \) and \( B \), where \( P(B) > 0 \), the probability of \( A \) given the condition \( B \) is

\[
P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}
\]

- The Multiplication Rule is

\[
P(A \text{ and } B) = P(A) \cdot P(B \mid A) \text{ or } P(B) \cdot P(A \mid B)
\]
Chapter 5 Probability Models

**Conditional Probability from the Sample Space**

P23. For the Titanic data in Display 5.39 on page 325, let $S$ be the event a person survived and $F$ be the event a person was female. Find and interpret these probabilities.

a. $P(F)$  
b. $P(F|S)$  
c. $P(\text{not } F)$

d. $P(\text{not } F|S)$  
e. $P(S|\text{not } F)$

P24. Display 5.44 gives the hourly workers in the United States, classified by race and by whether they were paid at or below minimum wage or above minimum wage. You select an hourly worker at random.

<table>
<thead>
<tr>
<th>Race</th>
<th>Paid at or Below Minimum Wage</th>
<th>Paid Above Minimum Wage</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>1,681</td>
<td>58,196</td>
<td>59,877</td>
</tr>
<tr>
<td>Black</td>
<td>227</td>
<td>9,190</td>
<td>9,417</td>
</tr>
<tr>
<td>Asian</td>
<td>38</td>
<td>2,634</td>
<td>2,672</td>
</tr>
<tr>
<td>Total</td>
<td>1,946</td>
<td>70,020</td>
<td>71,966</td>
</tr>
</tbody>
</table>


a. Find $P(\text{worker is paid at or below minimum wage}).$

b. Find $P(\text{worker is paid at or below minimum wage} \mid \text{worker is white}).$

c. What does a comparison of the two probabilities in parts a and b tell you?

d. Find $P(\text{worker is black}).$

e. Find $P(\text{worker is black} \mid \text{worker is paid at or below minimum wage}).$

f. What does a comparison of the two probabilities in parts d and e tell you?

P25. Suppose Jack draws marbles at random, without replacement, from a bag containing three red and two blue marbles. Find these conditional probabilities.

a. $P(\text{2nd draw is red} \mid \text{1st draw is red})$

b. $P(\text{2nd draw is red} \mid \text{1st draw is blue})$

c. $P(\text{3rd draw is blue} \mid \text{1st draw is red and 2nd draw is blue})$

d. $P(\text{3rd draw is red} \mid \text{1st draw is red and 2nd draw is red})$

P26. Suppose Jill draws a card from a standard 52-card deck. Find the probability that

a. it is a club, given that it is black

b. it is a jack, given that it is a heart

c. it is a heart, given that it is a jack

**The Multiplication Rule for $P(A \text{ and } B)$**

P27. Look again at the Titanic data in Display 5.39 on page 325.

a. Make a tree diagram to illustrate this situation, this time branching first on whether the person survived.

b. Write these probabilities as unreduced fractions.

i. $P(\text{survived})$

ii. $P(\text{female} \mid \text{survived})$

iii. $P(\text{survived and female})$

c. Now write a formula that tells how the three probabilities in part b are related. Compare it to the computation on page 328.

d. Write two formulas involving conditional probability to compute $P(\text{male and survived})$.

P28. Use the Multiplication Rule to find the probability that if you draw two cards from a deck without replacing the first before drawing the second, both cards will be hearts. What is the probability if you replace the first card before drawing the second?

P29. Suppose you take a random sample of size $n = 2$, without replacement, from the population {$W, W, M, M$}. Find these probabilities: $P(\text{W chosen 1st})$ and $P(\text{W chosen 2nd} \mid \text{W chosen 1st})$. Now find $P(\text{W chosen 1st and W chosen 2nd})$.

P30. Use the Multiplication Rule to find the probability of getting a sum of 8 and doubles when you roll two dice.
The Definition of Conditional Probability

P31. Suppose you roll two dice. Use the definition of conditional probability to find $P(\text{doubles} \mid \text{sum} = 8)$. Compare this probability with $P(\text{sum} = 8 \mid \text{doubles})$.

P32. Suppose you know that, in a class of 30 students, 10 students have blue eyes and 20 have brown eyes. Twenty-four of the students are right-handed, and 6 are left-handed. Of the left-handers, 2 have blue eyes. Make and fill in a table showing this situation. Then use the definition of conditional probability to find the probability that a student randomly selected from this class is right-handed, given that the student has brown eyes.

P33. As of July 1 of a recent season, the Los Angeles Dodgers had won 53% of their games. Eighteen percent of their games had been played against left-handed starting pitchers. The Dodgers won 36% of the games played against left-handed starting pitchers. What percentage of their games against right-handed starting pitchers did they win? [Source: Los Angeles Times, July 1, 2000, page D8.]

Conditional Probability and Medical Tests

P34. A laboratory technician is being tested on her ability to detect contaminated blood samples. Among 100 samples given to her, 20 are contaminated, each with about the same degree of contamination. Suppose the technician makes the correct decision 90% of the time.

a. Make a two-way table showing what you would expect to happen.
b. Compute and interpret her PPV.
c. Compute and interpret her NPV.
d. How would these rates change if she were given 100 samples with 50 contaminated?

P35. Display 5.45 presents hypothetical data on 130 people taking an inkblot test for bureaucratic pomposity disorder (BPD). Use the data to compute the four conditional probabilities defined for screening tests on page 331. Is the test a good one, in your judgment? Why or why not?

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Positive</th>
<th>Negative</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPD Present</td>
<td>60</td>
<td>40</td>
<td>100</td>
</tr>
<tr>
<td>Absent</td>
<td>20</td>
<td>10</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td>80</td>
<td>50</td>
<td>130</td>
</tr>
</tbody>
</table>

Display 5.45 Results of a test for bureaucratic pomposity disorder.

Conditional Probability and Statistical Inference

P36. Rosa has two pennies. One is an ordinary fair penny, but the other came from a magic shop and has two heads. Rosa chooses one of these coins, but she does not choose it randomly. She also doesn’t tell you how she chose it, so you don’t know the probabilities $P(\text{coin is two-headed})$ or $P(\text{coin is fair})$. She flips the coin once. For each probability, give a numerical value if it is possible to find one. If it is not possible to compute a probability, explain why.

a. $P(\text{coin lands heads} \mid \text{coin is fair})$
b. $P(\text{coin lands heads} \mid \text{coin is two-headed})$
c. $P(\text{coin is fair} \mid \text{coin lands heads})$
d. $P(\text{coin is two-headed} \mid \text{coin lands heads})$
e. $P(\text{coin is fair} \mid \text{coin lands tails})$
f. $P(\text{coin is two-headed} \mid \text{coin lands tails})$
Exercises

E41. Display 5.46 (at the bottom of this page) gives a breakdown of the U.S. population age 16 and older by age and whether the person volunteers time to charitable activities. You are working for a polling organization that is about to select a random sample of U.S. residents age 16 and older. Find these probabilities for the first person selected in the random sample.

a. \( P(\text{person is a volunteer}) \)

b. \( P(\text{person is a volunteer} | \text{person is age 16 to 24}) \)

c. \( P(\text{person is age 16 to 24} | \text{person is a volunteer}) \)

d. \( P(\text{person is age 16 to 34} | \text{person is not a volunteer}) \)

e. Which age group has the highest percentage of volunteers?

E42. Joseph Lister (1827–1912), British surgeon at the Glasgow Royal Infirmary, was one of the first to believe in Louis Pasteur’s germ theory of infection. He experimented with using carbolic acid to disinfect operating rooms during amputations. When carbolic acid was used, 34 of 40 patients lived. When carbolic acid was not used, 19 of 35 patients lived. Suppose a patient is selected at random. Make a two-way table and find

a. \( P(\text{patient died} | \text{carbolic acid used}) \)

b. \( P(\text{carbolic acid used} | \text{patient died}) \)

c. \( P(\text{carbolic acid used and patient died}) \)

d. \( P(\text{carbolic acid used or patient died}) \)

E43. A class collects information on students’ gender and handedness. The table gives the proportion of class members in each category.

a. If possible, fill in the rest of this table.

E44. Suppose the only information you have about the Titanic disaster appears in this two-way table. Is it possible to fill in the rest of the table? Is it possible to find the probability that a randomly selected person aboard the Titanic was female and survived? Either calculate the numbers, or explain why it is not possible to do so without additional information and tell what information you would need.

Display 5.46 Persons (in thousands) who performed unpaid volunteer activities in the last year. [Source: U.S. Census Bureau, Statistical Abstract of the United States, 2006, Table 575.]
E45. As you have seen, a useful way to work with the Multiplication Rule is to record the probabilities along the branches of a tree diagram. To practice this, suppose you draw two marbles at random, without replacement, from a bucket containing three red (R) and two blue (B) marbles. Compute the probabilities in parts a–f and then use these probabilities to fill in and label the branches in a copy of the tree diagram in Display 5.47.

a. \( P(R \text{ on 1st draw}) \)
b. \( P(B \text{ on 1st draw}) \)
c. \( P(R \text{ on 2nd draw} | R \text{ on 1st draw}) \)
d. \( P(B \text{ on 2nd draw} | R \text{ on 1st draw}) \)
e. \( P(R \text{ on 2nd draw} | B \text{ on 1st draw}) \)
f. \( P(B \text{ on 2nd draw} | B \text{ on 1st draw}) \)

Display 5.47 Tree diagram for drawing marbles.

E46. Your sock drawer hasn't been organized in a while. It contains six identical brown socks, three identical white socks, and five identical black socks. You draw two socks at random (without replacement).

a. Make a tree diagram similar to that in E45 to show the possible results of your draws.
b. Find \( P(\text{first sock is white}) \).
c. Find \( P(\text{second sock is white} | \text{first sock is white}) \).
d. Find \( P(\text{second sock is white} | \text{first sock is brown}) \).
e. What is the probability that you get two socks that match?

E47. Suppose you draw two cards from a 52-card deck.

a. What is the probability that both cards are aces if you replace the first card before drawing the second?
b. What is the probability that both cards are aces if you don't replace the first card before drawing the second?
c. What is the probability that the first card is an ace and the second is a king if you replace the first card before drawing the second?
d. What is the probability that the first card is an ace and the second is a king if you don't replace the first card before drawing the second?
e. What is the probability that both cards are the same suit if you replace the first card before drawing the second?
f. What is the probability that both cards are the same suit if you don't replace the first card before drawing the second?

E48. A Harris poll found that 36% of all adults said they would be interested in going to Mars. Of those who were interested in going, 62% were under age 25. Write a question for which the solution would be to compute \( (0.36)(0.62) \).

[Source: Harris, 1997.]

E49. Suppose that if rain is predicted, there is a 60% chance that it will actually rain. If rain is not predicted, there is a 20% chance that it will rain. Rain is predicted on 10% of days.

a. On what percentage of days does it rain?
b. Describe a simulation that you could run to check your answer.

E50. In 1999, Elizabeth Dole was a candidate to become the first woman president in U.S. history, and many observers assumed that she would show particular strength among women voters. According to a Gallup poll, “She did slightly better among Republican women than among Republican men, but this strength was not nearly enough to enable her to challenge Bush. In the October poll, Dole received the vote of 16% of Republican women, compared to 7%
of Republican men.” Given the additional information that the Republican party is about 60% male, find the probability that a Republican randomly selected from the October survey would have voted for Dole. [Source: www.gallup.com, 2000.]

E51. A laboratory screening test for the detection of a certain disease gives a positive result 6% of the time for people who do not have the disease. The test gives a negative result 0.5% of the time for people who do have the disease. Large-scale studies have shown that the disease occurs in about 3% of the population.

a. Fill in this two-way table, showing the results expected for every 100,000 people.

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Disease Yes</th>
<th>Disease No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3000</td>
<td>997000</td>
<td>100,000</td>
</tr>
</tbody>
</table>

b. What is the probability that a person selected at random tests positive for this disease?

c. What is the probability that a person selected at random who tests positive for the disease does not have the disease?

E52. A screening test for bureaucratic pomposity disorder (BPD) has reasonably good sensitivity and specificity, say 90% for both. (This means that, on average, nine out of every ten people who have BPD test positive and nine out of every ten who don't have BPD test negative.) Consider testing two different populations.

a. Officiousville has a population of 1000. Half the people in Officiousville have BPD. Suppose everyone in Officiousville gets tested. Fill out a table like Display 5.42 on page 330. Then use it to compute the PPV and NPV for this test.

b. Mellowville also has a population of 1000, but only 10 people in the whole town have BPD. Fill out a table to show how the test would perform if you used it to screen Mellowville. Then compute the PPV and NPV for this test.

E53. A new test for HIV, the virus that causes AIDS, uses a mouth swab instead of a blood test. The president of the company that produces this test was quoted as saying that overall the test was more than 99% accurate. The president claimed that, in the last year, the company received complaints of only 107 false positives out of 28,436 tests. [Source: New York Times, December 10, 2005.]

a. Can you determine the positive predictive value (PPV) for this test? Explain.

b. The article gave no information about false negatives. Why do you think no information was given?

c. How many false negatives would there have to be in order for the test to produce correct results in 99% of the 28,436 tests?

E54. Make a table and do a computation to illustrate that if the population being screened has a high percentage of people who have a certain disease, the NPV will tend to be low.

E55. Return to the situation in P36 in which Rosa has two pennies. One is an ordinary fair penny, but the other came from a magic shop and has two heads. Rosa chooses one of these coins, this time at random. She flips the coin once. For each probability, give a numerical value if it is possible to find one. If it is not possible, explain why not.

a. \( P(\text{coin lands heads} \mid \text{coin is fair}) \)

b. \( P(\text{coin lands heads} \mid \text{coin is two-headed}) \)

c. \( P(\text{coin is fair} \mid \text{coin lands heads}) \)

d. \( P(\text{coin is two-headed} \mid \text{coin lands heads}) \)

e. \( P(\text{coin is fair} \mid \text{coin lands tails}) \)

f. \( P(\text{coin is two-headed} \mid \text{coin lands tails}) \)

E56. One of these two statements is true, and one is false. Which is which? Make up an example using hypothetical data to illustrate your decision.

\[
P(A) = P(A \mid B) + P(A \mid \text{not } B)
\]

\[
P(A) = P(A \text{ and } B) + P(A \text{ and } \text{not } B)
\]
5.5 Independent Events

Jack and Jill have finished conducting taste tests with 100 adults from their neighborhood. They found that 60 of them correctly identified the tap water, but how well a person did depended on whether that person regularly uses bottled water for drinking. Of people who regularly drink bottled water, \( \frac{24}{30} \), or 80%, correctly identified the tap water. Of people who drink tap water, only \( \frac{36}{70} \), or 51.4%, correctly identified the tap water. (See Display 5.48.)

<table>
<thead>
<tr>
<th>Drinks Bottled Water?</th>
<th>Identified Tap Water?</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td></td>
<td>24</td>
<td>6</td>
<td>30</td>
</tr>
<tr>
<td>No</td>
<td></td>
<td>36</td>
<td>34</td>
<td>70</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>60</td>
<td>40</td>
<td>100</td>
</tr>
</tbody>
</table>

**Display 5.48** Results of taste test depend on whether a person regularly drinks bottled water.

On the other hand, as Display 5.49 shows, men and women did equally well in identifying the tap water. In each case, 60% correctly identified the tap water.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Identified Tap Water?</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td></td>
<td>21</td>
<td>14</td>
<td>35</td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td>39</td>
<td>26</td>
<td>65</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>60</td>
<td>40</td>
<td>100</td>
</tr>
</tbody>
</table>

**Display 5.49** Results of taste test don’t depend on gender.

Suppose one person is selected at random from the 100 who completed the taste test. Based on these data, the events *drinks bottled water* and *correctly identifies tap water* are dependent events. If the person drinks bottled water, he or she is more likely to correctly identify tap water. The events *is a male* and *correctly identifies tap water* are independent events. Being male doesn’t change the probability that the person correctly identifies tap water.

More formally, events \( A \) and \( B \) are independent if the probability of event \( A \) doesn’t depend on whether event \( B \) happens. That is, knowing that \( B \) happens doesn’t change the probability that \( A \) happens.

**Definition of Independent Events**

Suppose \( P(A) > 0 \) and \( P(B) > 0 \). Then events \( A \) and \( B \) are independent events if and only if \( P(A | B) = P(A) \) or, equivalently, \( P(B | A) = P(B) \).
In E73, you will show that \( P(A \mid B) = P(A) \) if and only if \( P(B \mid A) = P(B) \). In E74, you will show that if events \( A \) and \( B \) are independent, then so are events \( A \) and \( \text{not } B \) (event \( \text{not } B \) is the complement of event \( B \), consisting of all possible outcomes that are not in \( B \)).

**Example: Water, Gender, and Independence**

Show that the events *is a male* and *correctly identifies tap water* are independent and that the events *drinks bottled water* and *correctly identifies tap water* are not independent.

**Solution**

Suppose a person is selected at random from the 100 completing the water taste test. First, consider the events *is a male* and *correctly identifies tap water*.

\[
P(\text{correctly identifies tap water}) = \frac{60}{100} = 0.60
\]

\[
P(\text{correctly identifies tap water} \mid \text{is a male}) = \frac{21}{35} = 0.60
\]

Because these two probabilities are equal, the events are independent.

Next, consider the events *drinks bottled water* and *correctly identifies tap water*.

\[
P(\text{correctly identifies tap water}) = \frac{60}{100} = 0.60
\]

\[
P(\text{correctly identifies tap water} \mid \text{drinks bottled water}) = \frac{24}{30} = 0.80
\]

Because these two probabilities are not equal, the events are not independent.

**DISCUSSION**

**Independent Events**

D24. Suppose a person is selected at random from the 100 completing the water taste test. Show in another way that the events *is a male* and *correctly identifies tap water* are independent and that the events *drinks bottled water* and *correctly identifies tap water* are not independent. Specifically, use *correctly identifies tap water* as the condition in both cases.

D25. Suppose you choose a student at random from your school. In each case, does knowing that event \( A \) happened increase the probability of event \( B \), decrease the probability of event \( B \), or leave the probability of event \( B \) unchanged?

a. \( A: \) The student is a football player; \( B: \) The student weighs less than 120 lb.

b. \( A: \) The student has long fingernails; \( B: \) The student is female.

c. \( A: \) The student is a freshman; \( B: \) The student is male.

d. \( A: \) The student is a freshman; \( B: \) The student is a senior.
Multiplication Rule for Independent Events

In Section 5.4, you learned how to use the Multiplication Rule, \( P(A \text{ and } B) = P(A) \cdot P(B | A) \). If \( A \) and \( B \) are independent, then \( P(B | A) = P(B) \), so the Multiplication Rule reduces to a simple product.

Two events \( A \) and \( B \) are independent if and only if
\[
P(A \text{ and } B) = P(A) \cdot P(B)
\]

More generally, events \( A_1, A_2, \ldots, A_n \) are independent if and only if
\[
P(A_1 \text{ and } A_2 \text{ and } \ldots \text{ and } A_n) = P(A_1) \cdot P(A_2) \cdot \ldots \cdot P(A_n)
\]

You can also use this rule to decide whether two events are independent.

If \( P(A \text{ and } B) = P(A) \cdot P(B) \), then \( A \) and \( B \) are independent.

Example: Four Flips

If you flip a fair coin four times, what is the probability that all four flips result in heads?

Solution

The outcomes of the first flips don’t change the probabilities on the remaining flips, so the flips are independent. For a sequence of four flips of a fair coin,

\[
P(\text{heads on 1st flip and heads on 2nd and heads on 3rd and heads on 4th}) = P(\text{heads on 1st}) \cdot P(\text{heads on 2nd}) \cdot P(\text{heads on 3rd}) \cdot P(\text{heads on 4th})
\]

\[
= \left(\frac{1}{2}\right) \cdot \left(\frac{1}{2}\right) \cdot \left(\frac{1}{2}\right) \cdot \left(\frac{1}{2}\right) = \frac{1}{16}
\]

Example: Multiplication Rule and Health Insurance

According to Health Highlights, published by the U.S. Department of Health and Human Services, about 30% of young American adults ages 19 to 29 don’t have health insurance. [Source: www.healthfinder.gov, May 4, 2005.] What is the chance that if you choose two American adults from this age group at random, the first has health insurance and the second doesn’t?

Solution

Because the sample is selected from the large number of young adults in the United States, you can model the outcomes of the two trials as independent events. Then

\[
P(1st \ has \ insurance \ and \ 2nd \ doesn’t \ have \ insurance) = (0.3)(0.7) = 0.21
\]
You can conveniently list the probabilities of all four possible outcomes for two young adults in a "multiplication" table, as in Display 5.50.

<table>
<thead>
<tr>
<th></th>
<th>No Insurance</th>
<th>Insurance</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>First Young Adult</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Insurance</td>
<td>(0.3)(0.3) = 0.09</td>
<td>(0.3)(0.7) = 0.21</td>
<td><strong>0.30</strong></td>
</tr>
<tr>
<td>Insurance</td>
<td>(0.7)(0.3) = 0.21</td>
<td>(0.7)(0.7) = 0.49</td>
<td><strong>0.70</strong></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>0.30</strong></td>
<td><strong>0.70</strong></td>
<td><strong>1.00</strong></td>
</tr>
</tbody>
</table>

Display 5.50  Two-way table for the health insurance example.

**Example: Computing the Probability of "At Least One"**

About 30% of young American adults ages 19 to 29 don’t have health insurance. Suppose you take a random sample of ten American adults in this age group. What is the probability that at least one of them doesn’t have health insurance?

**Solution**

This question is asking for the probability that the first young adult doesn't have health insurance or the second young adult doesn't have health insurance or . . . or the tenth young adult doesn't have health insurance. You cannot use the Addition Rule for Mutually Exclusive Events and add 0.3 + 0.3 + · · · + 0.3, because the events aren't mutually exclusive (and note that you'll get a probability larger than 1). However, you can easily compute the complement, the probability that all the young adults have health insurance:

\[
P(\text{at least one doesn't have insurance})
= 1 - P(\text{all have insurance})
= 1 - P(1st has insurance \text{ and } 2nd has insurance \ldots \text{ and } 10th has insurance)
= 1 - P(1st has insurance) \cdot P(2nd has insurance) \cdot \cdots \cdot P(10th has insurance)
= 1 - (0.7)(0.7) \cdots \cdot (0.7)
= 1 - (0.7)^{10}
\approx 1 - 0.028
= 0.972
\]

You can use the Multiplication Rule because the very large number of young adults in the United States means that the probability of getting someone with insurance does not change (by any noticeable amount) depending on who has already been chosen.

In practice, computing \(P(A \text{ and } B)\) for independent events is straightforward: Use the definition, which tells you to multiply. What's not always straightforward in practice is recognizing whether independence gives a good model for a real situation. The basic idea is to ask whether one event has a bearing on the likelihood of another. This is something you try to decide by thinking about
the events in question. As always, models are simplifications of the real thing, and you’ll often use independence as a simplifying assumption to enable you to calculate probabilities, at least approximately.

**DISCUSSION**

**Multiplication Rule**

D26. Are these situations possible? Explain your answers.
   a. Events $A$ and $B$ are disjoint and independent.
   b. Events $A$ and $B$ are not disjoint and independent.
   c. Events $A$ and $B$ are disjoint and dependent.
   d. Events $A$ and $B$ are not disjoint and dependent.

D27. Suppose that you randomly and independently select three houses from those sold recently in the Chicago area. Find $P(\text{all three are below the median price})$. Find $P(\text{at least one is below the median price})$.

**Independence with Real Data**

You have seen that data often serve as a basis for establishing a probability model or for checking whether an assumed model is reasonable. In using data to check for independence, however, you have to be careful. For example, it would be fairly unusual for Jack and Jill to come up with exactly the same percentage of men as women who can identify tap water even if men and women are equally likely to be able to identify tap water. That example was “cooked up”; it’s almost impossible to find such perfection in practice.

**Example: Independence and Baseball**

In a recent season, by July 1 the Los Angeles Dodgers had won 41 games and lost 37 games. The breakdown by whether the game was played during the day or at night is shown in Display 5.51. If one of these games is chosen at random, are the events $\text{win}$ and $\text{day game}$ independent?

<table>
<thead>
<tr>
<th>Time of Game</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day</td>
<td>11</td>
<td>10</td>
<td>21</td>
</tr>
<tr>
<td>Night</td>
<td>30</td>
<td>27</td>
<td>57</td>
</tr>
<tr>
<td>Total</td>
<td>41</td>
<td>37</td>
<td>78</td>
</tr>
</tbody>
</table>

Display 5.51  Dodger record of wins and losses by time of game.  
[Source: Los Angeles Times, July 1, 2000, page D8.]

**Solution**

First, check the probabilities:

\[ P(\text{win}) = \frac{41}{78} = 0.526 \]

\[ P(\text{win} \mid \text{day game}) = \frac{11}{21} = 0.524 \]
Because these two probabilities aren’t exactly equal, you must conclude that the events \textit{win} and \textit{day game} are not independent. Yet, given that the Dodgers played 21 day games, the percentage of day games won couldn’t be any closer to 0.526. A statistician would conclude that if these results can be considered a random sample of their games, there isn’t sufficient evidence to say that the two events \textit{win} and \textit{day game} are not independent. This is the subject of Chapter 10.

Activity 5.5a further illustrates the difficulty of establishing independence with real data.

\section*{Independence with Real Data}

\textbf{What you’ll need:} one penny per student

1. Collect and study data on eyedness and handedness.
   a. Are you right-handed or left-handed?
   b. Determine whether you are right-eyed or left-eyed: Hold your hands together in front of you at arm's length. Make a space between your hands that you can see through. Through the space, look at an object at least 15 ft away. Now close your right eye. Can you still see the object? If so, you are left-eyed. Now close your left eye. Can you still see the object? If so, you are right-eyed.
   c. Would you expect being right-handed and right-eyed to be independent?
   d. Complete a two-way table for members of your class, showing the frequencies of eyedness and handedness.
   e. For a randomly selected student in your class, are being right-handed and right-eyed independent?

2. Collect and study data on the results of two coin flips.
   a. With other members of your class, flip a coin twice until you have 100 pairs of flips. Place your results in a two-way table. (Save this data for use again in Chapter 8.)
   b. Would you expect the results of the first flip and the second flip to be independent?
   c. From your data for two coin flips, do the results of the first flip and the second flip appear to be independent?

3. Did you get the results you expected in steps 1 and 2? Explain.

If a person is selected at random from the general population, being right-handed and being right-eyed aren’t independent, which is probably what you found for your class in Activity 5.5a. On the other hand, two flips of a coin are independent, even though for real data—actual flips—\(P(A)\) almost never exactly equals \(P(A \mid B)\), as the definition of independence requires.
In Chapter 10, you will learn how to examine tables like the ones for your coin flips in Activity 5.5a and for the Dodger record and then perform a statistical test of independence. The key question, just as in Chapter 1, will be whether the data are too extreme (too far from the strict definition of independence) to be consistent with that hypothesis.

**Summary 5.5: Independent Events**

The word *independent* is used in several related ways in statistics. First, two events are independent if knowledge that one event occurs does not affect the probability that the other event occurs. Second, in sampling units from a population with replacement, the outcome of the second selection is independent of what happened on the first selection (and the same is true for any other pair of selections). Sampling without replacement leads to dependent outcomes on successive selections. (However, if the population is large compared to the sample, you can compute probabilities as if the selections were independent without much error.) In both situations, the same definition of independence applies: Two events $A$ and $B$ are independent if and only if $P(A) = P(A | B)$; the probability of $A$ does not change with knowledge that $B$ has happened. If $P(A) = P(A | B)$, then it is also true that $P(B) = P(B | A)$.

You learned in this section that the Multiplication Rule simplifies to

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

if and only if $A$ and $B$ are independent events.

**Practice**

**Independent Events**

P37. Suppose you select one person at random from the *Titanic* passengers and crew in Display 5.39 on page 325. Use the definition of independent events to determine whether the events *didn't survive* and *male* are independent. Are any two events in this table independent?

P38. Suppose you draw a card at random from a standard deck. Use the definition of independent events to determine which pairs of events are independent.
   a. getting a heart; getting a jack
   b. getting a heart; getting a red card
   c. getting a 7; getting a heart

**Multiplication Rule for Independent Events**

P39. About 42% of people have type O blood. Suppose you select two people at random and check whether they have type O blood.
   a. Make a table like Display 5.50 on page 342 to show all possible results.
   b. What is the probability that exactly one of the two people has type O blood?
   c. Make a tree diagram that illustrates the situation.

P40. Suppose you select ten people at random. Using the information in P39, find the probability that
   a. at least one of them has type O blood
   b. at least one of them doesn't have type O blood

P41. After taking college placement tests, freshmen sometimes are required to repeat high school work. Such work is called “remediation” and does not count toward a college degree. About 11% of college freshmen have to take a remedial course in reading. Suppose you select two freshmen at
random and check to see if they have
to take remedial reading. [Source: C. Adelman,
Principal Indicators of Student Academic Histories in
Postsecondary Education, 1972–2000 (2004), Tables 7.1
and 7.2, preview.ed.gov. Data from U.S. Department of
Education, NCES, National Education Longitudinal Study

a. Find the probability that both freshmen
have to take remedial reading.

b. Show how to use a table like Display
5.50 on page 342 to find the probability
that exactly one freshman needs to take
remedial reading.

c. Show how to use a tree diagram to answer
the question in part b.

### Exercises

**E57.** Which of these pairs of events, $A$ and $B$, do you expect to be independent? Give a reason
for your answer.

a. For a test for tuberculosis antibodies:
   
   $A$: The test is positive.
   $B$: The person has a relative with tuberculosis.

b. For a test for tuberculosis antibodies:
   
   $A$: A person’s test is positive.
   $B$: The last digit of the person’s Social Security number is 3.

c. For a randomly chosen state in the United States:
   
   $A$: The state lies east of the Mississippi River.
   $B$: The state’s highest elevation is more than 8000 ft.

**E58.** Use the definition of independent events to determine which of these pairs of events are
independent when you roll two dice.

a. rolling doubles; rolling a sum of 8
b. rolling a sum of 8; getting a 2 on the first die rolled
c. rolling a sum of 7; getting a 1 on the first die rolled
d. rolling doubles; rolling a sum of 7
e. rolling a 1 on the first die; rolling a 1 on the second die

**E59.** Display 5.52 gives the handedness and eyedness of a randomly selected group of
100 people. Suppose you select a person from this group at random.

<table>
<thead>
<tr>
<th></th>
<th>Right-Eyed</th>
<th>Left-Eyed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right-Handed</td>
<td>57</td>
<td>31</td>
<td>88</td>
</tr>
<tr>
<td>Left-Handed</td>
<td>6</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>Total</td>
<td>63</td>
<td>37</td>
<td>100</td>
</tr>
</tbody>
</table>

**Display 5.52** Eyedness and handedness of 100 people.

a. Find each probability.
   
   i. $P($left-handed$)$
   ii. $P($left-eyed$)$
   iii. $P($left-eyed $| $left-handed$)$
   iv. $P($left-handed $| $left-eyed$)$

b. Are being left-handed and being left-eyed independent events?

c. Are being left-handed and being left-eyed mutually exclusive events?

**E60.** Display 5.53 gives the decisions on all applications to two of the largest graduate
programs at the University of California, Berkeley, by gender of the applicant. Suppose
you pick an applicant at random.

<table>
<thead>
<tr>
<th></th>
<th>Admit</th>
<th>Deny</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Man</td>
<td>650</td>
<td>592</td>
<td>1242</td>
</tr>
<tr>
<td>Woman</td>
<td>220</td>
<td>263</td>
<td>483</td>
</tr>
<tr>
<td>Total</td>
<td>870</td>
<td>855</td>
<td>1725</td>
</tr>
</tbody>
</table>

**Display 5.53** Application decisions for the two largest graduate programs at the
University of California, Berkeley, by gender. [Source: David Freedman, Robert
Pisani, and Roger Purves, Statistics, 3rd ed. (New York: Norton, 1997); data from Graduate Division,
UC Berkeley.]
a. Find each probability.
   i. \( P(\text{admitted}) \)
   ii. \( P(\text{admitted} \mid \text{woman}) \)
   iii. \( P(\text{admitted} \mid \text{man}) \)
b. Are being admitted and being a woman independent events?
c. Are being admitted and being a woman mutually exclusive events?
d. Is there evidence of possible discrimination against women applicants? Explain.
Display 5.54 gives the admissions data broken down by which program the applicant applied to.

<table>
<thead>
<tr>
<th></th>
<th>Admit</th>
<th>Deny</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Man</td>
<td>512</td>
<td>313</td>
<td>825</td>
</tr>
<tr>
<td>Woman</td>
<td>89</td>
<td>19</td>
<td>108</td>
</tr>
<tr>
<td>Total</td>
<td>601</td>
<td>332</td>
<td>933</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Admit</th>
<th>Deny</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Man</td>
<td>138</td>
<td>279</td>
<td>417</td>
</tr>
<tr>
<td>Woman</td>
<td>131</td>
<td>244</td>
<td>375</td>
</tr>
<tr>
<td>Total</td>
<td>269</td>
<td>523</td>
<td>792</td>
</tr>
</tbody>
</table>

Display 5.54 Application decisions for two of the largest graduate programs at the University of California, Berkeley, by program and gender.

e. Does Program A show evidence of possible discrimination against women applicants? Explain.
f. Does Program B show evidence of possible discrimination against women applicants? Explain.
g. These data illustrate Simpson's paradox. What is the paradox?

E61. When a baby is expected, there are two possible outcomes: a boy and a girl. However, they aren’t equally likely. About 51% of all babies born are boys.
   a. List the four possible outcomes for a family that has two babies (no twins).
   b. Which of the outcomes is most likely? Which is least likely?
   c. What is the probability that both children will be boys in a two-child family? What assumption are you making?

E62. Shondra has six shirts (blue, green, red, yellow, and two identical white shirts) and four pairs of pants (brown slacks, black slacks, and two identical pairs of jeans). Suppose Shondra selects a pair of pants at random and a shirt at random. Make a tree diagram showing Shondra’s choices. Use the tree diagram to compute the probability that she wears a white shirt with black pants.

E63. Forty-two percent of people in a town have type O blood, and 5% are Rh-positive. Assume blood type and Rh type are independent.
   a. Find the probability that a randomly selected person has type O blood and is Rh-positive.
   b. Find the probability that a randomly selected person has type O blood or is Rh-positive.
   c. Make a table that summarizes this situation.

E64. Schizophrenia affects 1% of the U.S. population and tends to first appear between the ages of 18 and 25. Today, schizophrenia accounts for the fifth highest number of years of work lost due to disability among Americans ages 15–44. Fewer than 30% of people with this illness are currently employed. Suppose you select a person from the U.S. population at random. State whether each question can be answered using the information given. If the question can be answered, answer it. [Source: National Institutes of Health, www.nih.gov]
   a. What is the probability that the person is schizophrenic?
b. What is the probability that the person is unemployed?

c. What is the probability that the person is unemployed and schizophrenic?

d. What is the probability that the person is unemployed or schizophrenic?

e. What is the probability that the person is unemployed given that he or she is schizophrenic?

f. What is the probability that the person is schizophrenic given that he or she is unemployed?

E65. A family has two girls and two boys. To select two children to do the dishes, the mother writes the names of the children on separate pieces of paper, mixes them up, and draws two different pieces of paper at random. What is the probability that both are girls?

E66. A committee of three students is to be formed from six juniors and five seniors. If three names are drawn at random, what is the probability that the committee will consist of all juniors or all seniors?

E67. As stated in P41, about 11% of college freshmen have to take a remedial course in reading. Suppose you take a random sample of 12 college freshmen.

a. What is the probability that all 12 will have to take a remedial course in reading?

b. What is the probability that at least one will have to take a remedial course in reading?

E68. As stated in E64, about 1% of the U.S. population are schizophrenic. Suppose you take a random sample of 50 U.S. residents.

a. What is the probability that none will be schizophrenic?

b. What is the probability that at least one will be schizophrenic?

E69. In the “Ask Marilyn” column of PARADE magazine, this question appeared:
Suppose a person was having two surgeries performed at the same time. If the chances of success for surgery A are 85% and the chances of success for surgery B are 90%, what are the chances that both would fail?

Write an answer to this question for Marilyn. [Source: PARADE, November 27, 1994, page 13.]

E70. Joaquin computes the probability that if two dice are rolled, then exactly one die will show a 4 as

\[ P(\text{4 on one die and not 4 on the other}) = \frac{\binom{5}{1}}{\binom{6}{1}} = \frac{5}{36} \]

Joaquin is wrong. What did he forget?

E71. A batter hits successfully 23% of the time against right-handed pitchers but only 15% of the time against left-handed pitchers. Suppose the batter faces right-handed pitchers in 80% of his at bats and left-handed pitchers in 20% of his at bats. What percentage of the time does he hit successfully?

E72. In Dave Barry Talks Back, Dave relates that he and his wife bought an Oriental rug “in a failed attempt to become tasteful.” Before going out to dinner, they admonished their dogs Earnest and Zippy to stay away from the rug.

“NO!!” we told them approximately 75 times while looking very stern and pointing at the rug. This proven training technique caused them to slink around the way dogs do when they feel tremendously guilty but have no idea why. Satisfied, we went out to dinner. I later figured out, using an electronic calculator, that this rug covers approximately 2 percent of the total square footage of our house, which means that if you (not you personally) were to have a random... [Dave’s word omitted because this is a textbook] attack in our home, the odds are approximately 49 to 1 against your having it on our Oriental rug. The odds against your having four random attacks on this rug are more than five million to one. So we had to conclude that it was done on purpose.
Do you agree with Dave’s arithmetic? Do you agree with his logic? [Source: Dave Barry, Dave Barry Talks Back, © 1991 by Dave Barry. Used by permission of Crown Publishers, a division of Random House, Inc.]

E73. Two events \(A\) and \(B\) are independent if any of these statements holds true:

\[
P(A) = P(A|B)
\]

\[
P(B) = P(B|A)
\]

\[
P(A \text{ and } B) = P(A) \cdot P(B)
\]

Use the definition of conditional probability to show that these three conditions are equivalent as long as \(P(A)\) and \(P(B)\) don’t equal 0. That is, show that if any one is true, so are the other two.

E74. Show that if events \(A\) and \(B\) are independent, then events \(A\) and \(\text{not } B\) are also independent.

E75. \(P(A) = P(A|B) \cdot P(B) + P(A|\overline{B}) \cdot P(\overline{B})\) sometimes is called the law of total probability.

a. Prove the law of total probability.

b. Use this law to find \(P(\text{Dodgers win})\) from the information in Display 5.51 on page 343.

c. Prove that

\[
P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A|B) \cdot P(B) + P(A|\overline{B}) \cdot P(\overline{B})}
\]

This result is called Bayes’ Theorem.

d. Use Bayes’ Theorem to find \(P(\text{Dodgers win | day game})\).

E76. Two screening tests are used on patients chosen from a certain population. Test I, the less expensive of the two, is used about 60% of the time and produces false positives in about 10% of cases. Test II produces false positives in only about 5% of cases.

a. Fill in this table of percentages.

<table>
<thead>
<tr>
<th>Results</th>
<th>Test I</th>
<th>Test II</th>
<th>Total %</th>
</tr>
</thead>
<tbody>
<tr>
<td>False Positives</td>
<td>—?—</td>
<td>—?—</td>
<td>—?—</td>
</tr>
<tr>
<td>Other Results</td>
<td>—?—</td>
<td>—?—</td>
<td>—?—</td>
</tr>
<tr>
<td>Total %</td>
<td>60</td>
<td>40</td>
<td>100</td>
</tr>
</tbody>
</table>

b. A false positive is known to have occurred in a patient tested by one of these two tests. Find the approximate probability that Test I was used, first by direct observation from the table and then by use of the formula in E75, part c.

Chapter Summary

Probability is the study of random behavior. The probabilities used in statistical investigations often come from a model based on equally likely outcomes. In other cases, they come from a model based on observed data. While you do not know for sure what the next outcome will be, a model based on many observations gives you a pretty good idea of the possible outcomes and their probabilities.

Two important concepts were introduced in this chapter: disjoint (mutually exclusive) events and independent events.

- Two events are disjoint if they can’t happen on the same opportunity. If \(A\) and \(B\) are disjoint events, then \(P(A \text{ and } B) = 0\).
- Two events are independent if the occurrence of one doesn’t change the probability that the other will happen. That is, events \(A\) and \(B\) are independent if and only if \(P(A) = P(A|B)\).
You should know how to use each of these rules for computing probabilities:

- The probability that event $A$ doesn't happen is
  \[ P(\text{not } A) = 1 - P(A) \]
  This rule often is used to find the probability of at least one success:
  \[ P(\text{at least one success}) = 1 - P(\text{no successes}) \]

- The Addition Rule gives the probability that event $A$ or event $B$ or both occur:
  \[ P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \]
  If the events $A$ and $B$ are disjoint, this rule simplifies to
  \[ P(A \text{ or } B) = P(A) + P(B) \]

- The Multiplication Rule gives the probability that events $A$ and $B$ both occur:
  \[ P(A \text{ and } B) = P(A) \cdot P(B | A) \text{ or } P(B) \cdot P(A | B) \]
  If $A$ and $B$ are independent events, this rule simplifies to
  \[ P(A \text{ and } B) = P(A) \cdot P(B) \]

### Review Exercises

E77. Suppose you roll a fair four-sided (tetrahedral) die and a fair six-sided die.

a. How many equally likely outcomes are there?

b. Show all the outcomes in a table or in a tree diagram.

c. What is the probability of getting doubles?

d. What is the probability of getting a sum of 3?

e. Are the events getting doubles and getting a sum of 4 disjoint? Are they independent?

f. Are the events getting a 2 on the tetrahedral die and getting a 5 on the six-sided die disjoint? Are they independent?

E78. Backgammon is one of the world’s oldest games. Players move counters around the board in a race to get “home” first. The number of spaces moved is determined by a roll of two dice. If your counter is able to land on your opponent’s counter, your opponent has to move his or her counter to the “bar” where it is trapped until your opponent can free it. For example, suppose your opponent’s counter is five spaces ahead of yours. You can “hit” that counter by rolling a sum of 5 with both dice or by getting a 5 on either die.

a. Use the sample space for rolling a pair of dice to find the probability of being able to hit your opponent’s counter on your next roll if his or her counter is five spaces ahead of yours.

b. Can you use the Addition Rule for Disjoint Events to compute the probability that you roll a sum of 5 with both dice or get a 5 on either die? Either do the computation or explain why you can’t.

E79. Jorge has a CD player attached to his alarm clock. He has set the CD player so that when it’s time for him to wake up, it randomly selects one song to play on the CD. Suppose
there are nine songs on the CD and Jorge’s favorite song is the third one.

a. Describe a simulation to estimate the solution to this probability problem: What is the probability that Jorge will hear his favorite song at least once in the next week?

b. Compute this probability exactly.

E80. This problem appeared in the “Ask Marilyn” column of PARADE magazine and stirred up a lot of controversy. Marilyn devoted at least four columns to it.

A woman and a man (unrelated) each have two children. At least one of the woman’s children is a boy, and the man’s older child is a boy. Do the chances that the woman has two boys equal the chances that the man has two boys? [Source: PARADE, July 27, 1997.]

a. Design a simulation to answer this question. (The process of designing a simulation will help you clarify your assumptions.)

b. What is the answer?

E81. As you saw in the example on page 305, about 75% of men and 90% of women wash their hands after using a public restroom. Suppose you observe one man and one woman at random. Compute each of these probabilities.

a. \( P(\text{both the man and the woman wash their hands}) \)

b. \( P(\text{neither washes his or her hands}) \)

c. \( P(\text{at least one washes his or her hands}) \)

d. \( P(\text{man washes his hands} | \text{woman washes her hands}) \)

E82. As you saw in E18 on page 312, about 25% of U.S. residents claim to believe in astrology. Suppose you select two U.S. residents at random. Compute each of these probabilities.

a. \( P(\text{both believe in astrology}) \)

b. \( P(\text{neither believes in astrology}) \)

c. \( P(\text{at least one believes in astrology}) \)

d. \( P(\text{the second person believes in astrology} | \text{the first person believes in astrology}) \)

E83. Among recent graduates of mathematics departments, half intend to teach high school. A random sample of size 2 is to be selected from the population of recent graduates.

a. If mathematics departments had only four recent graduates total, what is the chance that the sample will consist of two graduates who intend to teach high school?

b. If mathematics departments had 10 million recent graduates, what is the chance that the sample will consist of two graduates who intend to teach high school?

c. Are the selections technically independent in part a? Are they technically independent in part b? In which part can you assume independence anyway? Why?

E84. If you select a student at random from the school described, what is the probability that the student has ridden a merry-go-round or a roller coaster, given the information about that school? Make a table to illustrate each situation.

a. In a particular school, 80% of students have ridden a merry-go-round and 45% have ridden a roller coaster. Only 15% have done neither.

b. In another school, 30% of students have ridden a merry-go-round but not a roller coaster. Forty-five percent have ridden a roller coaster but not a merry-go-round. Only 20% have done neither.

E85. A website at Central Michigan University collects data about college students. Students were asked their political party preference and which candidate they voted for in the last presidential election. Display 5.55 (on the next page) shows the results for the first 42 students who were 22 years old or older and had either Democrat or Republican party preference. Suppose you pick one of these 42 students at random.
Display 5.55 Results of voting and political party preference survey. [Source: Central Michigan University, stat.cst.cmich.edu.]

<table>
<thead>
<tr>
<th>Party Preference</th>
<th>Voted for Democrat</th>
<th>Voted for Republican</th>
<th>Did Not Vote</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democrat</td>
<td>19</td>
<td>1</td>
<td>6</td>
<td>26</td>
</tr>
<tr>
<td>Republican</td>
<td>1</td>
<td>12</td>
<td>3</td>
<td>16</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>13</td>
<td>9</td>
<td>42</td>
</tr>
</tbody>
</table>

a. Find each of these probabilities.
- i. $P(\text{Republican})$
- ii. $P(\text{voted Republican})$
- iii. $P(\text{voted Republican and Republican})$
- iv. $P(\text{voted Republican or Republican})$
- v. $P(\text{Republican} | \text{voted Republican})$
- vi. $P(\text{voted Republican} | \text{Republican})$

b. Are being Republican and voting Republican independent events?

c. Are being Republican and voting Democrat mutually exclusive events?

E86. Display 5.56 gives probabilities for a randomly selected adult in the United States.

<table>
<thead>
<tr>
<th>Age</th>
<th>18 to 25</th>
<th>25 or Older</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>$p$</td>
<td>$q$</td>
<td>$p + q$</td>
</tr>
<tr>
<td>No</td>
<td>$r$</td>
<td>$s$</td>
<td>$r + s$</td>
</tr>
<tr>
<td>Total</td>
<td>$p + r$</td>
<td>$q + s$</td>
<td>1</td>
</tr>
</tbody>
</table>

Display 5.56 Cell phone ownership, by age.

Write an expression for the probability that a randomly selected adult
- a. has a cell phone
- b. is 25 or older, given that the person has a cell phone
- c. has a cell phone, given that the person is age 18 to 25
- d. is 25 or older and has a cell phone
- e. is 25 or older or has a cell phone

E87. Consider a situation with these three probabilities:

$$P(A) = \frac{1}{4}, \quad P(B | A) = \frac{1}{2}, \quad P(B | \text{not A}) = \frac{1}{4}$$

a. Fill in a two-way table that has $A$ and not $A$ for one way and $B$ and not $B$ for the other way.

b. Are events $A$ and $B$ independent?

E88. Jack and Jill found that 60 out of 100 people correctly identified the tap water. They are uncertain what to conclude in their report. They are concerned that perhaps everyone was simply guessing and that 60 of them got the right answer simply by chance. So Jack and Jill flipped a fair coin 100 times and counted the number of heads. They did this over and over again. Out of thousands of trials, they got 60 or more heads about 2.5% of the time. Write a conclusion for their report.

E89. Many people who have had a stroke get depressed. These patients tend to have poorer outcomes than stroke patients who aren't depressed. A standard, but lengthy, test for depression was given to 79 stroke patients and 43 were found to be clinically depressed. Researchers wondered if a single question, “Do you often feel sad or depressed?” would be just as good as the longer test. Of the 43 patients found to be clinically depressed, 37 answered “yes,” and 6 answered “no.” Of the 36 patients classified as not clinically depressed, 8 answered “yes” and 28 answered “no.” [Source: C. Watkins et al., Accuracy of a Single Question in Screening for Depression in a Cohort of Patients After Stroke: Comparative Study. BMJ 2001;323:1159 (17 November).]

Does the single question appear to be a good diagnostic test for clinical depression in stroke patients? First make a table that summarizes this situation. Then use the ideas of sensitivity, specificity, positive predictive value, and negative predictive value in your answer.
E90. A telephone area code is a three-digit number that begins with a digit from 2 through 9. The second digit can be any number from 1 through 8 or 0. The third number must be different from the second. How many area codes are possible, according to these rules?

E91. A few years ago, the U.S. Post Office added the “zip plus 4” to the original five-digit zip codes. For example, the zip code for the national headquarters of the American Red Cross is now 20006-5310.

For this problem, assume that a zip code can be any five-digit number except 00000 and that the four-digit “plus 4” can be any number except 0000.

a. Assume that the U.S. population is roughly 300 million. If all possible five-digit zip codes are used, what is the average number of people per zip code?

b. If all possible “zip plus 4” codes are used, what is the average number of people per “zip plus 4”?

E92. DNA contains coded information that tells living organisms how to build proteins by combining 20 different amino acids. The code is based on the sequence of building blocks called nucleotides. There are four kinds of nucleotides grouped into triplets called codons. Each codon corresponds to an amino acid. In a codon, order doesn't matter, and the same nucleotide can be used more than once. So, for example, AAB is the same as ABA. Show that there is exactly one codon for each of the 20 amino acids.
AP1. A student argues that extraterrestrials will either abduct her statistics teacher by tomorrow or they will not, and therefore there’s a 1 out of 2 chance for each of these two events. Which of the following best explains why this reasoning is incorrect?

A. The two events are not independent.
B. The two events are mutually exclusive.
C. The two events are not equally probable.
D. The two events are complements.
E. There are more than two events that need to be considered.

AP2. Suppose you roll two dice. Which of the following are independent events?

A. getting a sum of 8; getting doubles
B. getting a sum of 3; getting doubles
C. getting a sum of 2; getting doubles
D. getting a 1 on the first die; getting a sum of 5
E. getting a 1 on the first die; getting doubles

AP3. Suppose you roll two dice. Which of the following are mutually exclusive (disjoint) events?

A. getting a sum of 8; getting doubles
B. getting a sum of 3; getting doubles
C. getting a sum of 2; getting doubles
D. getting a 1 on the first die; getting a sum of 5
E. getting a 1 on the first die; getting doubles

AP4. Suppose 15 children visit a retirement home at various times during one day, and each child randomly chooses one of 10 retirees to visit. A “successful” day is one in which all 10 retirees are visited by at least one of these children. A statistics student wants to use simulation to estimate the probability of a successful day. How should the student conduct one run?

A. Assign the digits 0–9 to the retirees.
   Assign the integers 1–15 to the children. Randomly choose a number from each group, pairing up a child and a retiree. Repeat, without replacement, until all 10 retirees are assigned a child. Record whether the day is successful.
B. Assign the integers 0–15 to the children. Randomly choose digits one at a time, with replacement, until 10 different integers are chosen. Record the number of integers needed to get 10 different ones.
C. Assign the digits 0–9 to the retirees. Randomly choose 15 digits, with replacement, and record whether all 10 digits were chosen or not.
D. Assign the integers 1–15 to the children. Randomly choose 10 of these integers, without replacement, and record the proportion that are less than 10.
E. Assign the integers 1–15 to the children. Randomly choose 10 of these integers, with replacement, and record whether all 10 integers were different or not.

AP5. In a statistics classroom, 50% of the students are female and 30% of the students got an A on the most recent test. What is the probability that a student picked at random from this classroom is a female who got an A on the most recent test?

A. 0.15
B. 0.20
C. 0.40
D. 0.65
E. cannot be determined from the information given
AP6. In the fictional country of Valorim, 4% of adults are smokers who get lung cancer, 8% are nonsmokers who get lung cancer, 22% are smokers who do not get lung cancer, and the remaining 66% are nonsmokers who do not get lung cancer. What is the probability that a randomly selected adult smoker gets lung cancer?

A 4/100  
B 4/30  
C 4/26  
D 4/22  
E 4/12

AP7. Seventy percent of Valorites can answer the previous question correctly. Suppose you select three Valorites at random. What is the probability that at least one of the three can answer correctly?

A 0.100  
B 0.343  
C 0.900  
D 0.973  
E 2.100

AP8. A new space probe, the Mars Crashlander, has a main antenna and a backup antenna. The probe will arrive at Mars and crash into the surface. The main antenna has a 50% chance of surviving the crash. The backup has a 90% chance of surviving if the main antenna survives, and a 20% chance of surviving if the main antenna fails. What is the probability that the Crashlander will have at least one working antenna after the crash?

A 10%  
B 27.5%  
C 55%  
D 60%  
E 70%

Investigative Tasks

AP9. “So they stuck them down below, where they’d be the first to go.” This line from the song about the Titanic refers to the third-class passengers. The first two-way table in Display 5.57 gives the number of surviving Titanic passengers in various classes of travel. The members of the ship’s crew have been omitted from this table. The second and third tables break down survival by class and by gender. Use the ideas about conditional probability and independence that you have learned to analyze these tables.

<table>
<thead>
<tr>
<th>Class</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>203</td>
</tr>
<tr>
<td>Second</td>
<td>118</td>
</tr>
<tr>
<td>Third</td>
<td>178</td>
</tr>
<tr>
<td>Total</td>
<td>499</td>
</tr>
<tr>
<td>Survived?</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>325</td>
</tr>
<tr>
<td>No</td>
<td>817</td>
</tr>
<tr>
<td>Total</td>
<td>1316</td>
</tr>
</tbody>
</table>

Females: Class

<table>
<thead>
<tr>
<th>Class</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>141</td>
</tr>
<tr>
<td>Second</td>
<td>93</td>
</tr>
<tr>
<td>Third</td>
<td>90</td>
</tr>
<tr>
<td>Total</td>
<td>324</td>
</tr>
<tr>
<td>Survived?</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>145</td>
</tr>
<tr>
<td>No</td>
<td>123</td>
</tr>
<tr>
<td>Total</td>
<td>447</td>
</tr>
</tbody>
</table>

Males: Class

<table>
<thead>
<tr>
<th>Class</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>62</td>
</tr>
<tr>
<td>Second</td>
<td>25</td>
</tr>
<tr>
<td>Third</td>
<td>88</td>
</tr>
<tr>
<td>Total</td>
<td>175</td>
</tr>
<tr>
<td>Survived?</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>180</td>
</tr>
<tr>
<td>No</td>
<td>510</td>
</tr>
<tr>
<td>Total</td>
<td>694</td>
</tr>
</tbody>
</table>

Display 5.57 Titanic survival data, by class.  
[Source: Journal of Statistics Education 3, no. 3 (1995).]
Forty percent of people have type A blood. A blood bank is in dire need of a type A donor today. How many donors will they have to test before finding the first type A? A probability distribution can describe the chances of the possible outcomes.
A **probability distribution** describes the possible numerical outcomes of a chance process and allows you to find the probability of any set of possible outcomes. Sometimes a probability distribution is defined by a table, like the ones Jack and Jill made in Chapter 5. Sometimes a probability distribution is defined by a curve, like the normal curve in Chapter 2. If you select a male at random, all possible heights he could have are given by the values on the $x$-axis, and the probability of getting someone whose height is between two specified $x$-values is given by the area under the curve between those two $x$-values. As you will learn in this chapter, sometimes a probability distribution is defined by a formula.

Probability distributions for practical use come about in two different ways—through data collection and through theory. If you want to know the chance of a paper cup landing on its closed bottom when tossed, the best way to find out is to toss the cup many times and collect data. On the other hand, if you want the probability of an event that can be modeled by coin flipping, you can use the fact that the probability of heads is always 0.5 and build from there. This chapter begins with data collection but quickly moves to theory.

Some types of probability distributions occur so frequently in practice that it is important to know their names and formulas. Among these are the binomial and geometric distributions, which closely (but not perfectly) reflect many real-world situations and so are used to model many applications that have similar characteristics.

**In this chapter, you will learn**

- the terminology of probability distributions
- to construct probability distributions from data or theory
- the concept and uses of expected value
- to recognize and apply the binomial distribution
- to recognize and apply the geometric distribution
Random Variables and Expected Value

In Chapter 4 you learned about sampling, and in Chapter 5 you learned about probability. What is the connection between these two topics?

- **Sampling** deals with estimating a characteristic of a population by selecting data from that population. A random sample of households from your community helps you estimate the percentage of all households in your community that own a motor vehicle.

- **Probability** deals with the chance of getting a specified outcome in a sample when you know about the population. If you know 91% of households in your community own a motor vehicle, then you can calculate the probability that three randomly selected households all own a motor vehicle.

One of the most important questions statistics tries to answer is “What might an unknown population look like?” A census, which would answer this question, is usually impossible or impractical. Fortunately, early 20th-century statisticians had this extraordinary insight: Use probability to learn what samples from populations with known distributions tend to look like. Then, take a sample from the unknown population and compare the results to those from known populations. The populations for which your sample appears to fit right in are the plausible homes for your sample. This chapter begins to formalize this profound idea, which will be developed more fully in the remaining chapters of the book.

**Probability Distributions from Data**

Suppose you are working for a contractor who is building 500 new single-family houses and wants to know how many parking spaces will be needed per house. Your job is to sample 500 households for the purpose of estimating the number of vehicles per household. Fortunately, before starting the work of taking a survey, you find a distribution of the number of vehicles per household (Display 6.1) from the U.S. Census Bureau. Assuming that these national results are typical of your community, your work is done! You can report to your boss, for example, that about 0.385 + 0.137 + 0.058, or 0.580 (more than half) of the households will have two vehicles or more.

Your boss is pleased that you have so much extra time and gives you a new project. Your construction company is also building 500 duplexes, and your boss wants to know something about the total number of vehicles that will be parked by the two households occupying a duplex.

Display 6.1 does not give you that information. You are beginning to think you will have to take a survey when you realize that you can take two households at random from the distribution in Display 6.1 and add their numbers. Repeating this 500 times, you will get an approximation to the distribution your boss is asking for. This approximation depends on the assumptions that households living in duplexes mirror households in general with respect to vehicles per household, and that the number of vehicles in one household is independent of the number of vehicles in the neighboring household. How good the approximation is depends on how well these assumptions hold up. This is typical of all modeling problems; the accuracy of the result depends on the appropriateness of the model.
You begin by cutting up 1000 identical slips of paper, planning to write 0 on 88 of them, 1 on 332 of them, 2 on 385 of them, 3 on 137 of them, and 4 on 58 of them. You will then place the slips in a box, mix them up, and draw one at random. You will have an $\frac{88}{1000}$, or 0.088, chance of getting a household with no motor vehicles, a $\frac{332}{1000}$, or 0.332, chance of getting a household with one motor vehicle, and so on. You will replace the slip, mix the slips up, and draw another. Then you will add the two numbers to get the total number of vehicles for the first duplex. You will need 500 pairs of draws to simulate the results for 500 duplexes.

After cutting 321 slips of paper, a quicker method occurs to you—use a table of random digits! To use a list of random digits to select a household at random from this distribution, you make the assignments as in Display 6.2. You need three-digit numbers because the percentages have three decimal places.

For example, 88 out of 1000 (0.088) three-digit sequences are needed to represent households with no motor vehicles. Here the 88 numbers 001 through 088 were used, although you could use 000 through 087 or any other choice as long as there are 88 numbers. The next 332 three-digit sequences (089–420) represent households with one motor vehicle, and so on. Now that you’ve assigned all the
three-digit sequences, you’re ready to draw your pairs of households. Divide a string of random digits into groups of three.

3 9 1 | 5 4 5 | 1 7 7 | 3 2 4 | 1 0 6 | 8 4 5 | 2 4 8

The first sequence, 391, represents a household with one motor vehicle because 391 is in the interval 089–420. The second sequence, 545, represents a household with two motor vehicles. So for the first duplex, you have a total of three motor vehicles. In sampling this way, you might come across a sequence that you have already used. You should go ahead and use it again because each random sequence represents many households, not just one, and you want to keep the probabilities the same for each household selected.

Such a simulation for 500 duplexes (each consisting of randomly selecting two values from the distribution in Display 6.1 and adding them) is shown in Display 6.3. This distribution differs considerably from the one in Display 6.1 and could not be anticipated without some clever work with probabilities.

<table>
<thead>
<tr>
<th>Total Number of Vehicles</th>
<th>Proportion of Duplexes</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.008</td>
</tr>
<tr>
<td>1</td>
<td>0.058</td>
</tr>
<tr>
<td>2</td>
<td>0.142</td>
</tr>
<tr>
<td>3</td>
<td>0.306</td>
</tr>
<tr>
<td>4</td>
<td>0.250</td>
</tr>
<tr>
<td>5</td>
<td>0.160</td>
</tr>
<tr>
<td>6</td>
<td>0.064</td>
</tr>
<tr>
<td>7</td>
<td>0.010</td>
</tr>
<tr>
<td>8</td>
<td>0.002</td>
</tr>
</tbody>
</table>

Display 6.3 Simulated distribution of total vehicles in two households.

You can now report to your boss, for example, that there is an 0.088 chance that a single-family household will have no vehicles but only an 0.008 chance that the occupants of a duplex will have no vehicles.

**DISCUSSION**

**Probability Distributions from Data**

D1. When you sample from distributions like the one in Display 6.1 using random digits, explain why repeated random digits should be used again rather than discarded.

D2. Compare the shape, center, and spread of the distribution in Display 6.3 with that in Display 6.1.

D3. Your boss finds it odd that there is a fairly good chance (0.137 + 0.058, or 0.195) that a single household will need to park three or more vehicles while there is little chance (0.064 + 0.010 + 0.002, or 0.076) that the two households in a duplex will need to park six or more vehicles. Explain to your boss why this is reasonable.
Probability Distributions from Theory

The language of probability distributions reflects the chance involved in random outcomes. The variable of interest, $X$, is called a **random variable** because its numerical values vary depending on the results of a particular trial. Thus, if $X$ represents the number of motor vehicles in a randomly selected household, $P(X = 2) = 0.385$, because about 38.5% of randomly selected households have two motor vehicles.

You now have two ways to establish probability distributions from data. Sometimes relative frequencies from data already collected can serve as a probability distribution for a random variable to be studied in the future, as with the distribution of motor vehicles per household from the Census Bureau. At other times a simulation is a good way to build an approximate probability distribution, as with the approximation of the probability distribution of the total number of vehicles in two randomly selected households in a duplex. There is a third way: Use the rules of theoretical probability to construct a probability distribution from basic principles and assumptions. The next three examples illustrate this approach.

**Example: Probability Distributions from Rolling Two Dice**

In Chapter 5, you constructed a table like Display 6.4, which lists all possible outcomes when two dice are rolled. From this table, you can construct different probability distributions, depending on the numerical summary that interests you.

<table>
<thead>
<tr>
<th>Second Die</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1, 1</td>
<td>1, 2</td>
<td>1, 3</td>
<td>1, 4</td>
<td>1, 5</td>
<td>1, 6</td>
</tr>
<tr>
<td>2</td>
<td>2, 1</td>
<td>2, 2</td>
<td>2, 3</td>
<td>2, 4</td>
<td>2, 5</td>
<td>2, 6</td>
</tr>
<tr>
<td>3</td>
<td>3, 1</td>
<td>3, 2</td>
<td>3, 3</td>
<td>3, 4</td>
<td>3, 5</td>
<td>3, 6</td>
</tr>
<tr>
<td>4</td>
<td>4, 1</td>
<td>4, 2</td>
<td>4, 3</td>
<td>4, 4</td>
<td>4, 5</td>
<td>4, 6</td>
</tr>
<tr>
<td>5</td>
<td>5, 1</td>
<td>5, 2</td>
<td>5, 3</td>
<td>5, 4</td>
<td>5, 5</td>
<td>5, 6</td>
</tr>
<tr>
<td>6</td>
<td>6, 1</td>
<td>6, 2</td>
<td>6, 3</td>
<td>6, 4</td>
<td>6, 5</td>
<td>6, 6</td>
</tr>
</tbody>
</table>

Display 6.4  The 36 outcomes when rolling two dice.

For example, if you are playing Monopoly, you need to find the sum of the two dice. If you are playing backgammon, you might be more interested in the larger of the two numbers. Construct each of these probability distributions. Then find the probability that, when you roll two dice, the sum is 3. Then find the probability that the larger number is 3.

**Solution**

Display 6.5 shows these two probability distributions, one of the sum of two dice and one of the larger number on the two dice. (In the case of doubles, the larger number and the smaller number are the same.)
When you roll two dice, the probability that the sum of the numbers is 3 is \( \frac{2}{36} \). The probability that the larger number is 3 is \( \frac{5}{36} \).

**Example: Sampling Lung Cancer Patients**

Display 6.6 shows data on the proportion of lung cancer cases caused by smoking. Suppose two lung cancer patients are randomly selected from the large population of people with that disease. Construct the probability distribution of \( X \), the number of patients with lung cancer caused by smoking.

<table>
<thead>
<tr>
<th>Lung Cancer Cases</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoking responsible</td>
<td>0.9</td>
</tr>
<tr>
<td>Smoking not responsible</td>
<td>0.1</td>
</tr>
</tbody>
</table>

**Solution**

There are four possible outcomes for the two patients. With “yes” representing “caused by smoking” and “no” representing “not caused by smoking,” the possibilities are

- no for 1st patient and no for 2nd patient
- no for 1st patient and yes for 2nd patient
- yes for 1st patient and no for 2nd patient
- yes for 1st patient and yes for 2nd patient
The fact that the patients are selected at random from a large population implies that the outcomes for the two selections may be considered independent. Using the rules of Chapter 5,

\[ P(\text{no for 1st patient and no for 2nd patient}) = P(\text{no for 1st}) \cdot P(\text{no for 2nd}) \]
\[ = (0.1)(0.1) = 0.01 \]

\[ P(\text{no for 1st patient and yes for 2nd patient}) = (0.1)(0.9) = 0.09 \]
\[ P(\text{yes for 1st patient and no for 2nd patient}) = (0.9)(0.1) = 0.09 \]
\[ P(\text{yes for 1st patient and yes for 2nd patient}) = (0.9)(0.9) = 0.81 \]

The first of these outcomes results in \( X = 0 \), the second and third each result in \( X = 1 \), and the fourth results in \( X = 2 \). Because the second and third outcomes are disjoint, their probabilities can be added. The probability distribution of \( X \) is then given by this table:

<table>
<thead>
<tr>
<th>( X )</th>
<th>Probability of ( X )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.01</td>
</tr>
<tr>
<td>1</td>
<td>0.09 + 0.09 = 0.18</td>
</tr>
<tr>
<td>2</td>
<td>0.81</td>
</tr>
</tbody>
</table>

**Example: Sampling Hockey Teams**

Display 6.7 gives the capacity of the home arenas of the five teams in the Atlantic Division of the National Hockey League. Suppose you pick two teams at random to play a pair of exhibition games, with one game played in each home arena. Construct the probability distribution of \( X \), the total number of people who could attend the two games. Then find the probability that the total possible attendance will be at least 36,000.

<table>
<thead>
<tr>
<th>Team</th>
<th>Arena Seating (in thousands)</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Jersey Devils</td>
<td>19</td>
</tr>
<tr>
<td>New York Islanders</td>
<td>16</td>
</tr>
<tr>
<td>New York Rangers</td>
<td>18</td>
</tr>
<tr>
<td>Philadelphia Flyers</td>
<td>18</td>
</tr>
<tr>
<td>Pittsburgh Penguins</td>
<td>17</td>
</tr>
</tbody>
</table>

**Display 6.7** Home arena capacities.

**Solution**

There are ten ways to pick two teams from the five: \( \binom{5}{2} = \frac{5!}{2!(5-2)!} = 10 \). These ten pairs of teams, with the total possible attendance, are shown in Display 6.8.
The probability distribution of the total possible attendance for two games is given in Display 6.9.

<table>
<thead>
<tr>
<th>Total Possible Attendance, x</th>
<th>Probability, p</th>
</tr>
</thead>
<tbody>
<tr>
<td>33</td>
<td>0.1</td>
</tr>
<tr>
<td>34</td>
<td>0.2</td>
</tr>
<tr>
<td>35</td>
<td>0.3</td>
</tr>
<tr>
<td>36</td>
<td>0.2</td>
</tr>
<tr>
<td>37</td>
<td>0.2</td>
</tr>
</tbody>
</table>

The probability that the total attendance is at least 36,000 is 0.2 + 0.2, or 0.4.

Probability Distributions from Theory

D4. The boss of your construction company isn’t quite satisfied with the simulation in Display 6.3 on page 360 and would like to know the exact probability that two randomly selected duplex households have a total of two vehicles. Use the rules of probability from Chapter 5 and the information in Display 6.1 on page 359 to compute this probability exactly. How does your exact probability compare to the estimate in Display 6.3?

Expected Value and Standard Deviation

Your boss has looked at the table in Display 6.1 on page 359 and would like to know the average number of vehicles that would need parking spaces for a single-family house.

Assuming that your community is like the entire United States, you can find the mean number of vehicles per household directly from Display 6.1. You expect
8.8% of households to have no vehicles, 33.2% to have one vehicle, and so on. Thus, you expect the mean number of vehicles per household to be

$$\mu_X = 0(0.088) + 1(0.332) + 2(0.385) + 3(0.137) + 4(0.058) = 1.745$$

Computing the mean can be easily accomplished by adding a third column to the table, as in Display 6.10. Notice that the mean, 1.745, is the balance point of the histogram in Display 6.1.

<table>
<thead>
<tr>
<th>Vehicles per Household, x</th>
<th>Probability of Household, p</th>
<th>Contribution to Mean, x (\cdot) p</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.088</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0.332</td>
<td>0.332</td>
</tr>
<tr>
<td>2</td>
<td>0.385</td>
<td>0.770</td>
</tr>
<tr>
<td>3</td>
<td>0.137</td>
<td>0.411</td>
</tr>
<tr>
<td>4</td>
<td>0.058</td>
<td>0.232</td>
</tr>
<tr>
<td><strong>Sum</strong></td>
<td></td>
<td><strong>1.745</strong></td>
</tr>
</tbody>
</table>

Display 6.10  Computing the mean from a probability distribution.

The mean of a probability distribution has a special name, although continuing to call it the “mean” is not wrong.

The mean of a probability distribution for the random variable \(X\) is called its **expected value** and is usually denoted by \(\mu_X\) or \(E(X)\).

You can report to your boss that the expected number of vehicles per household is 1.745. However, you realize that you should give your boss an estimate of how much a typical household is likely to differ from this average. To calculate the standard deviation of the number of vehicles per household, you find the expected value of the square of the deviations from the mean, which is called the **variance** of the probability distribution. As always, the standard deviation is then the square root of the variance. The variance of the distribution in Display 6.1 on page 359 is given by

$$\sigma_X^2 = (0 - 1.745)^2(0.088) + (1 - 1.745)^2(0.332) + (2 - 1.745)^2(0.385)$$

$$+ (3 - 1.745)^2(0.137) + (4 - 1.745)^2(0.058) \approx 0.9880$$

The standard deviation is \(\sigma_X = \sqrt{0.9880} \approx 0.994\). You tell your boss that the neighborhood of single-family houses can expect to have 1.745 motor vehicles per house, but quite often a house will have about one vehicle more or less than this mean.

The formulas for the expected value (mean) and standard deviation of a probability distribution listed in a table, shown in the following box, parallel those for a relative frequency table.
Formulas for the Mean and Variance of a Probability Distribution Listed in a Table

The expected value (mean) $\mu_X$ and variance $\sigma_X^2$ of a probability distribution listed in a table are

$$E(X) = \mu_X = \sum x_i p_i \quad \text{and} \quad \text{Var}(X) = \sigma_X^2 = \sum (x_i - \mu_X)^2 p_i$$

where $p_i$ is the probability that the random variable $X$ takes on the specific value $x_i$. To get the standard deviation, take the square root of the variance.

Now your boss asks about the duplexes. For the duplexes, the simulated data of Display 6.3 on page 360 should provide a good approximation of the actual probability distribution of this random variable, call it $Y$. The mean, variance, and standard deviation as calculated from the 500 simulated values in the frequency table of Display 6.3 turn out to be

$$\mu_Y \approx 3.530 \quad \sigma_Y^2 \approx 1.8411 \quad \sigma_Y \approx 1.357$$

The neighborhood of duplexes can expect to have 3.530 motor vehicles per duplex but often will see up to 1.357 vehicles more or less than this.

Example: CDs Purchased by USC Students

The probability distribution in Display 6.11 was constructed from a survey at the University of Southern California (USC). The random variable is the number of CDs purchased by a student in the previous year. For example, if you select a USC student at random, the probability is 0.06 that he or she purchased no CDs in the previous 12 months. Estimate the expected value and standard deviation of this distribution.

### Display 6.11 CDs purchased by USC students.

<table>
<thead>
<tr>
<th>Number of CDs Purchased Within 12 Months, $x$</th>
<th>Probability, $p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.06</td>
</tr>
<tr>
<td>1–5</td>
<td>0.28</td>
</tr>
<tr>
<td>6–10</td>
<td>0.25</td>
</tr>
<tr>
<td>11–20</td>
<td>0.21</td>
</tr>
<tr>
<td>21–50</td>
<td>0.17</td>
</tr>
<tr>
<td>Over 50</td>
<td>0.03</td>
</tr>
</tbody>
</table>
| **Total**                                    | **1.00**         

Solution

**Student:** I remember how to do this. First I have to make a reasonable estimate of the center of each interval. For example, for the
interval of 1–5 CDs, I’ll use 3. Then I simply use the formulas:

\[
\mu_X = \sum x_i p_i \\
= 0(0.06) + 3(0.28) + 8(0.25) + 15.5(0.21) \\
+ 35.5(0.17) + 55(0.03) \\
= 13.78
\]

and

\[
\sigma_X^2 = \sum (x_i - \mu_X)^2 p_i \\
= (0 - 13.78)^2(0.06) + (3 - 13.78)^2(0.28) \\
+ (8 - 13.78)^2(0.25) + (15.5 - 13.78)^2(0.21) \\
+ (35.5 - 13.78)^2(0.17) + (55 - 13.78)^2(0.03) \\
= 184.0766
\]

So the standard deviation, \(\sigma_X\), is approximately 13.57.

**Statistician:** Your computations are perfect. Now, what is the expected number of CDs purchased by a USC student?

**Student:** Easy. My computations give 13.78, but because we can’t have part of a CD, we say the expected number of CDs purchased by a USC student is 13 or 14.

**Statistician:** And you were doing so well. It’s true you can’t have part of a CD, but expected value is the same thing as the average, and you know that the average doesn’t have to be one of the values in the distribution. So you should say that the expected number of CDs is 13.78. What does the fact that the mean and standard deviation are about equal tell you about the shape of this distribution?

**Student:** Well, the number of purchases can’t go below zero, which is only one standard deviation below the mean. So, there must be a long tail toward the larger values.

**Statistician:** You are rapidly becoming a statistical thinker!

You can use a calculator to quickly find the mean and variance of a probability distribution listed in a table. The mean and variance for the data in Display 6.11 are shown here. [See Calculator Note 6A to learn how to calculate these statistics.]
Expected Value and Standard Deviation

D5. In Chapter 2, you used this formula to compute the mean, $\bar{x}$, of the values in a frequency table, where $x$ is the value, $f$ is its frequency, and $n = \sum f$:

$$\bar{x} = \frac{\sum x \cdot f}{n}$$

a. Use this formula to compute the mean of the values in this frequency table.

<table>
<thead>
<tr>
<th>Value, $x$</th>
<th>Frequency, $f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>6</td>
<td>23</td>
</tr>
<tr>
<td>8</td>
<td>15</td>
</tr>
</tbody>
</table>

b. Suppose the data had been given in a relative frequency table like this one, which shows the proportion of times each value occurs. Fill in the rest of the second column.

<table>
<thead>
<tr>
<th>Value, $x$</th>
<th>Proportion, $f/n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.24</td>
</tr>
<tr>
<td>6</td>
<td>—?—</td>
</tr>
<tr>
<td>8</td>
<td>—?—</td>
</tr>
</tbody>
</table>

c. Show that you can find the mean, $\bar{x}$, using the formula

$$\sum \left( x \cdot \frac{f}{n} \right)$$

d. Discuss how the formula in part c relates to the formula for the expected value, $\mu$.

D6. Compare the mean number of vehicles in single-family households to the mean number in a duplex. Compare the variances. What do you notice?

D7. Eighteen percent of high school boys and 10% of high school girls say they rarely or never wear seat belts. Suppose one high school boy and one high school girl are selected at random, with the random variable of interest being the number in the pair who say they rarely or never wear seat belts. Describe two ways of finding the expected value and standard deviation of this random variable, at least approximately. (Use only the material that has been presented in this chapter so far.)

D8. Define a random variable for Display 6.4 on page 361 that is different from the two random variables in Display 6.5. Give its probability distribution and compute its expected value and standard deviation.

D9. This sentence appeared in the British humor magazine *Punch*:

The figure of 2.2 children per adult female was felt to be in some respects absurd, and a Royal Commission suggested that the middle classes be paid money to increase the average to a rounder and more convenient number.


Who would find the figure absurd—the student or the statistician from the dialogue on page 367?
Expected Value in Everyday Situations

In many real-life situations, you can determine the best course of action by considering expected value.

Example: The Wisconsin Lottery

The Wisconsin lottery had a scratch-off game called “Big Cat Cash.” It cost $1 to play. The probabilities of winning various amounts are listed in Display 6.12. Find and interpret the expected value when playing one game.

<table>
<thead>
<tr>
<th>Winnings, x</th>
<th>Probability, p</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1</td>
<td>1/10</td>
</tr>
<tr>
<td>$2</td>
<td>1/14</td>
</tr>
<tr>
<td>$3</td>
<td>1/24</td>
</tr>
<tr>
<td>$18</td>
<td>1/200</td>
</tr>
<tr>
<td>$50</td>
<td>1/389</td>
</tr>
<tr>
<td>$150</td>
<td>1/20,000</td>
</tr>
<tr>
<td>$900</td>
<td>1/120,000</td>
</tr>
</tbody>
</table>

Display 6.12  Probabilities for Wisconsin scratch-off game.
(Source: www.wilottery.com.)

Solution

First, note that the probabilities don’t sum to 1—it’s not even close. That’s because the most likely outcome is winning nothing. So imagine another row with $0 for “Winnings” and 0.7793 for “Probability.” Using the expected-value formula, you can verify that the expected value for this probability distribution is 0.6014. This means that if you spend $1 to play this game, you “expect” to get 60.14¢ back in winnings. Of course, you can’t get this amount on any one play, but over the long run that would be the average return. Another way to understand this is to imagine playing the game 1000 times. You expect to get back $601.40, but you will have spent $1000. The standard deviation of the winnings per game is $4.040, which is quite large because of the possibility of winning one of the larger amounts.

The expected value may not be of much importance to an individual player (unless he or she is going to play many times) but it is of great importance to the State of Wisconsin, which can expect to pay out $601.40 for every $1,000 bet.

Example: Burglary Insurance

A person who lives in a large city comes to your insurance agency and asks you to insure her household so that if it is burgled, you would pay her $5000. What should you charge per year for such insurance?

Solution

How much you charge per year depends on how likely it is that her household will be burgled. Insurance companies and the mathematicians they employ, called actuaries, keep careful records of various crimes and disasters so they can
know the probabilities that these will occur. Unfortunately, you neglected to hire an actuary, and all the information you can find is the nationwide rate of 41.9 burglaries per 1000 urban households.

Display 6.13 gives the two possible outcomes, their payouts, and their probabilities.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Payout</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>No burglary</td>
<td>$0</td>
<td>0.9581</td>
</tr>
<tr>
<td>Burglary</td>
<td>$5000</td>
<td>0.0419</td>
</tr>
</tbody>
</table>


The expected payout per policy is

\[ E(X) = \mu_X = 0(0.9581) + 5000(0.0419) = 209.50 \]

You expect to break even if you charge $209.50 for the insurance. But as a good businessperson, you will actually charge more to cover your costs of doing business and to give yourself some profit.

### Expected Value in Everyday Situations

D10. A few years ago, a fast-food restaurant chain had a scratch-off card promotion that contained four separate games. The player could play only one of them and could choose which one. The probability of winning game A was \( \frac{1}{2} \), and the prize was a drink worth $0.55. The probability of winning game B was \( \frac{1}{4} \), and the prize was a food item worth $0.69. The probability of winning game C was \( \frac{1}{8} \), and the food prize was worth $1.44. The probability of winning game D was \( \frac{1}{16} \), and the food prize was worth $1.99. Assuming you like all the prizes, which is the best game to play? This question is open to different interpretations, so we’ll be specific: On which game are your expected winnings greatest?

You can organize the information for each game in a table. For example, for game A, the table looks like this:

<table>
<thead>
<tr>
<th>Value of Prize</th>
<th>Probability of Winning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0.55</td>
<td>0.5</td>
</tr>
<tr>
<td>$0.00</td>
<td>0.5</td>
</tr>
</tbody>
</table>

The expected winnings if you play game A are

\[ \mu_X = (0.55)(0.5) + (0)(0.5) = 0.275 \]

or $0.275.

a. Compute the expected winnings from games B, C, and D. On which game are your expected winnings greatest?

b. For what reasons would people choose to play the other games, even if they knew the expected values were lower?
### Rules for Means and Variances

In Section 2.4, you learned that if you recenter a data set—that is, add the same number \( c \) to all the values—it doesn’t change the shape or standard deviation but adds \( c \) to the mean. If you rescale a data set—that is, multiply all the values by the same nonzero number \( d \)—it doesn’t change the basic shape but multiplies the mean by \( d \) and the standard deviation by \( |d| \) where \( |d| \) indicates the absolute value of \( d \). These rules for distributions of data also apply to probability distributions.

#### Linear Transformation Rule: The Effect of a Linear Transformation of \( X \) on \( \mu_X \) and \( \sigma_X \)

Suppose you have a probability distribution with random variable \( X \), mean \( \mu_X \) and standard deviation \( \sigma_X \). If you transform each value by multiplying it by \( d \) and then adding \( c \), where \( c \) and \( d \) are constants, then the mean and standard deviation of the transformed values are given by

\[
\mu_{c+dx} = c + d\mu_X \\
\sigma_{c+dx} = |d|\sigma_X
\]

#### Example: Tripling the Wisconsin Lottery

Suppose that, for a special promotion, the prizes are tripled for the Wisconsin scratch-off game “Big Cat Cash” described in Display 6.12 on page 369. What are the expected winnings for a person who plays one game? What is the standard deviation?

**Solution**

The most efficient method of doing this problem is to transform the expected value, 0.6014, and the standard deviation, 4.040, for the original game using the rules of linear transformation. Here, \( c = 0 \) and \( d = 3 \), so the expected value of the tripled prize is 3(0.6014), or $1.804, with a standard deviation of 3(4.040), or $12.12.

A second method of doing this problem is to rewrite Display 6.12 showing the winnings during the special promotion, as shown in Display 6.14.

<table>
<thead>
<tr>
<th>Old Winnings, ( x )</th>
<th>Winnings in Special Promotion, ( 3x )</th>
<th>Probability, ( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0</td>
<td>$0</td>
<td>0.7793</td>
</tr>
<tr>
<td>$1</td>
<td>$3</td>
<td>1/10</td>
</tr>
<tr>
<td>$2</td>
<td>$6</td>
<td>1/14</td>
</tr>
<tr>
<td>$3</td>
<td>$9</td>
<td>1/24</td>
</tr>
<tr>
<td>$18</td>
<td>$54</td>
<td>1/200</td>
</tr>
<tr>
<td>$50</td>
<td>$150</td>
<td>1/389</td>
</tr>
<tr>
<td>$150</td>
<td>$450</td>
<td>1/20,000</td>
</tr>
<tr>
<td>$900</td>
<td>$2700</td>
<td>1/120,000</td>
</tr>
</tbody>
</table>

Display 6.14 Winnings during special promotion.
Then use the formulas for the expected value and variance from page 366 as before:

\[
E(X) = 0(0.7793) + 3(1/10) + 6(1/14) + 9(1/24) + 54(1/200) + 150(1/389) \\
+ 450(1/20000) + 2700(1/120000) \\
= 1.804
\]

\[
\text{Var}(X) = (0 - 1.804)^2 \cdot 0.7793 + (3 - 1.804)^2(1/10) + (6 - 1.804)^2(1/14) \\
+ (9 - 1.804)^2(1/24) + (54 - 1.804)^2(1/200) \\
+ (150 - 1.804)^2(1/389) + (450 - 1.804)^2(1/20000) \\
+ (2700 - 1.804)^2(1/120000) \\
= 146.89
\]

\[
\sigma(x) = \sqrt{146.89} = 12.12
\]

The expected winnings are $1.804, with a standard deviation of $12.12.

Suppose that, instead of tripling the winnings, the special promotion is to give a person three tickets instead of just one. What are the mean and standard deviation of the winnings now? The rules in the next box will help you solve this new problem.

### Addition and Subtraction Rules for Random Variables

If \(X\) and \(Y\) are random variables, then

\[
\mu_{X+Y} = \mu_X + \mu_Y \\
\mu_{X-Y} = \mu_X - \mu_Y
\]

and if \(X\) and \(Y\) are independent, then

\[
\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 \\
\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2
\]

**Example: Getting Three Lottery Tickets**

Refer to the example about the Wisconsin lottery game with an expected value of $0.6014 and standard deviation $4.040. Suppose a person is given three tickets. What are the expected total winnings and the standard deviation of this total?
Solution

Let $X + Y + Z$ represent the person’s total winnings on three tickets. Then $\mu_{X+Y+Z}$ is the expected value of the total winnings on three tickets. This problem requires a generalization of the Addition Rule:

\[
\mu_{X+Y+Z} = \mu_X + \mu_Y + \mu_Z
\]
\[
= 0.6014 + 0.6014 + 0.6014
\]
\[
= 1.804
\]

Because the winnings on the three tickets can be considered independent, the variance of $X + Y + Z$ is

\[
\sigma^2_{X+Y+Z} = \sigma^2_X + \sigma^2_Y + \sigma^2_Z
\]
\[
= 4.04^2 + 4.04^2 + 4.04^2
\]
\[
\approx 48.965
\]

The standard deviation is

\[
\sigma = \sqrt{48.965} \approx 6.997
\]

The person expects winnings of $1.804 with standard deviation $6.997.

Note the difference in the previous two examples. In the first, there was one randomly selected ticket and its value was tripled. In the second, there were three randomly selected tickets and their values were added. The expected values are the same, but there is more variability when the winnings from a single ticket are tripled. The next example combines both sets of rules.

**Example: Expected Savings**

Suppose you earn $12 an hour for tutoring but spend $8 an hour for dance lessons. You save the difference between what you earn and the cost of your lessons. The number of hours you spend on each activity in a week varies independently according to the probability distributions in Display 6.15. Find your expected weekly savings and the standard deviation of your weekly savings.

<table>
<thead>
<tr>
<th>Hours of Dance Lessons, $x$</th>
<th>Probability, $p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.4</td>
</tr>
<tr>
<td>1</td>
<td>0.3</td>
</tr>
<tr>
<td>2</td>
<td>0.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hours of Tutoring, $y$</th>
<th>Probability, $p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.3</td>
</tr>
<tr>
<td>2</td>
<td>0.3</td>
</tr>
<tr>
<td>3</td>
<td>0.2</td>
</tr>
<tr>
<td>4</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Display 6.15  Probability distributions of the number of hours per week taking dance lessons and tutoring.

Solution

Let $X$ be the number of hours per week taking dance lessons and $Y$ be the number of hours per week tutoring. The expected number of hours you take dance lessons, $\mu_X$, is $0(0.4) + 1(0.3) + 2(0.3)$, or 0.9, with a standard deviation, $\sigma_X$, of
0.831 hour. Similarly, the expected number of hours you tutor, or $\mu_Y$, is 2.3, with a standard deviation, $\sigma_Y$, of 1.1 hours. You spend $8 for each hour of dance lessons, so by the Linear Transformation Rule your mean weekly expenditure for dance lessons is

$$\mu_{8X} = 8 \cdot \mu_X = (8) \cdot (0.9) = 7.20$$

or $7.20. Similarly, if you earn $12 for each hour of tutoring, then by the Linear Transformation Rule, your mean weekly earnings from tutoring are

$$\mu_{12Y} = 12 \cdot \mu_Y = (12) \cdot (2.3) = 27.60$$

or $27.60. By the Subtraction Rule, your expected weekly savings are

$$\mu_{12Y} - \mu_{8X} = 27.60 - 7.20 = 20.40$$

or $20.40.

By the Linear Transformation Rule, the standard deviation of the cost of your weekly lessons, $\sigma_{8X}$, is 8(0.831), or $6.648$, and the standard deviation of your weekly earnings from tutoring, $\sigma_{12Y}$, is 12(1.1), or $13.20$. Because the amounts are independent, you can now use the Subtraction Rule to find that the variance of your weekly savings is

$$\sigma^2_{12Y-8X} = \sigma^2_{12Y} + \sigma^2_{8X} = 13.2^2 + 6.648^2 = 218.435904$$

Taking the square root, the standard deviation, $\sigma_{12Y-8X}$, is about $14.78$.

For most students, the most surprising rule says that to get the variance of the difference of two independently selected variables, you add the individual variances. Why add and not subtract? Activity 6.1a will help you see the reason.

### ACTIVITY 6.1a

**The Sum and Difference of Two Rolls of a Die**

**What you’ll need:** a die, two partners

You will roll your die twice. One partner will record the sum of the numbers on the two rolls. The other partner will record the difference—the number on the first roll minus the number on the second roll. For example, if the first roll is 3 and the second roll is 5, the first partner records 8 and the second records −2.

1. What do you think the distribution of sums (first roll plus second roll) and the distribution of differences (first roll minus second roll) will look like after many rolls? Where will they be centered? Which will have the larger spread?
2. Roll the die until your class has recorded 100 sums and 100 differences.
3. Plot the 100 sums and the 100 differences on separate dot plots. How do the shapes compare? How do the spreads compare?
4. What is the mean of your distribution of sums? Of differences? How do these means compare to the mean of the distribution of a single roll of a die?

(continued)
5. Compute the variance of your distribution of sums. Compute the variance of your distribution of differences. How do the variances of your distributions of sums and differences compare with the variance of the outcomes from a single roll of a die? How do they compare with each other?

6. Use the formulas in the box on page 366 to compute the mean and variance of the distribution of sums. Use the same formulas to compute the mean and variance of the distribution of differences. How close are your results in step 5 to those you get using the formulas?

7. Explain why it should be the case that the spread in the differences is equal to the spread in the sums.

**Rules for Means and Variances**

D11. If you expect to work 10 hours next week with a standard deviation of 2 hours and you expect to study 15 hours next week with a standard deviation of 3 hours, what is the total number of hours you expect to spend working or studying? Use the rule for adding variances to estimate the standard deviation of the total number of hours spent studying or working. Is it reasonable to use the rule in this case? Explain.

D12. Provide an intuitive argument as to why the variability in a sum should be larger than the variability in either variable making up the sum. Do the same for the variability in a difference.

**Summary 6.1: Random Variables and Expected Value**

A probability distribution describes the possible numerical outcomes, $x$, of a chance process and allows you to find the probability, $P(x)$, of any set of possible outcomes. A probability distribution table for a given random variable lists each of these outcomes in one column and their associated probability in another column. You can compute the expected value (mean) and variance of a probability distribution table using the formulas

$$E(X) = \mu_X = \sum x_i p_i$$

$$Var(X) = \sigma^2_X = \sum (x_i - \mu_X)^2 p_i$$

Estimating the expected value has many real-world applications. For example, you can figure out your expected weekly savings or the break-even price for insurance.

For random variables $X$ and $Y$ and constants $c$ and $d$,

- the mean and standard deviation of a linear transformation of $X$ are given by

$$\mu_{c+dx} = c + d\mu_X$$

$$\sigma_{c+dx} = |d|\sigma_X$$
• if you add or subtract random variables $X$ and $Y$, the mean is given by

$$\mu_{X+Y} = \mu_X + \mu_Y$$

$$\mu_{X-Y} = \mu_X - \mu_Y$$

• if $X$ and $Y$ are independent, then the variance is given by

$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$$

$$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$$

Practice

Probability Distributions from Data

P1. Refer to Display 6.2 on page 359. Use this line of random digits to simulate selecting two households at random and counting the number of motor vehicles in both households together. Then repeat for two more households.

1 7 7 3 2 4 1 0 6 8 4 5 2 4 8

P2. As you saw in the example on page 362, 90% of lung cancer cases are caused by smoking. How would you assign the numbers in a random digit table so that they represent the distribution in Display 6.6? Use a random digit table (Table D on page 828) to select a lung cancer patient at random and then tell whether smoking was responsible for the patient’s lung cancer.

P3. The distribution in Display 6.16 gives the number of children per family in the United States. Describe how to use this line from a random digit table to find the number of children in a randomly selected family:

4 8 8 3 0 9 9 4 2 5 1 7 7 3 8 9 0 8 1 7

Use your process to find the total number of children in three randomly selected families.

<table>
<thead>
<tr>
<th>Number of Children</th>
<th>Proportion of Families</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.524</td>
</tr>
<tr>
<td>1</td>
<td>0.201</td>
</tr>
<tr>
<td>2</td>
<td>0.179</td>
</tr>
<tr>
<td>3</td>
<td>0.070</td>
</tr>
<tr>
<td>4 or more</td>
<td>0.026</td>
</tr>
</tbody>
</table>

Display 6.16 The number of children per family.

P4. Use the appropriate rule of probability from Chapter 5 and the fact that there is an 0.088 chance that a single household will have no vehicles to compute the probability that the two households in a duplex will have a total of zero vehicles. How close is your computation to the estimate in Display 6.3 on page 360?
**Probability Distributions from Theory**

P5. Suppose you roll two dice. Construct the probability distribution of the smaller of the two numbers.

P6. The best estimate of the percentage of all email that is spam (electronic junk mail) is about 70%. Suppose you randomly select three email messages. Construct the probability distribution of the random variable, \( X \), defined as the number of messages in your sample that are spam. Begin by listing the outcomes, as was done in the lung cancer example on page 362.

P7. Suppose five students (Claire, Charlotte, Max, Alisa, and Shaun) belong to a club and two of them (Alisa and Shaun) take a bus to school. You randomly select three students from the club to serve on a committee that meets before school. Your goal is to construct the probability distribution of the random variable, \( X \), defined as the number of committee members who take a bus to school.

a. List all possible random samples of size 3 from this group of five students.

b. Recalling that in simple random sampling all samples of the same size have equal probability of selection, assign a probability to each sample listed in part a.

c. Based on parts a and b, construct the probability distribution of the random variable \( X \).

**Expected Value in Everyday Situations**

P11. For a raffle, 500 tickets will be sold.

a. What is the expected value of a ticket if the only prize is worth $600?

b. What is the expected value of a ticket if there is one prize worth $1000 and two prizes worth $400 each?

P12. Refer to the example on burglary insurance on page 369. The burglary rate for a suburban location is 23.2 burglaries per 1000 households. If the home is in a suburban area, what should you charge for $5000 of burglary insurance in order to break even?

P13. A neighborhood hardware store rents a steam cleaner for 3 hours at a fee of $30.00 or for a day (up to 8 hours) at a fee of $50.00. Sometimes the cleaner can be rented out for two 3-hour periods in the same day. By studying rental records, the manager of the store comes up with approximate probabilities for renting the cleaner on a typical Saturday, as shown in Display 6.17 (on the next page).
### Rules for Means and Variances

**P14.** Construct a probability distribution for the difference when a die is rolled twice (first roll minus second roll).

a. What is the shape of this distribution?

b. Compute the mean and the variance of this distribution using the standard formulas

\[ \mu_X = \sum x_i p_i \quad \text{and} \quad \sigma^2_X = \sum (x_i - \mu_X)^2 p_i \]

c. Verify that the mean of this probability distribution is equal to \( \mu - \mu \), or 0, and the variance is \( \sigma^2 + \sigma^2 \), or \( 2\sigma^2 \), where \( \mu \) and \( \sigma \) are the mean and standard deviation of the distribution of a single roll of a die.

**P15.** Refer to the example on expected savings on page 373. Find the amount you expect to save in a year and the standard deviation of your yearly savings.

**P16.** Your construction company plans to build apartment buildings, each with one common parking lot. Your boss would like an estimate of how many spaces would be needed for a building with 200 households.

a. Using the mean number of vehicles per household, 1.745, from page 365, find how many vehicles you would expect for an apartment building with 200 households.

b. You realize that a randomly selected group of 200 households probably won't have exactly the expected number of vehicles. How far off is your estimate likely to be?

c. What assumptions are you making in your answers to parts a and b? Do you think they are reasonable for apartment dwellers?

**P17.** Find the expected value and standard deviation of the total number of vehicles per duplex using only the information from Display 6.1 on page 359. How do your answers compare with the approximate values found from Display 6.3 (recorded on page 360)?

**P18.** Display 6.18 shows the distribution of the number of color televisions per household in the United States.

![Color televisions per household](source: U.S. Census Bureau, Statistical Abstract of the United States, 2006.)

a. A random variable is defined as the number of color televisions observed in a single household selected at random from across the country. Find the (approximate) expected value and standard deviation for the probability distribution of this random variable. Mark the expected value on a plot of this distribution.

b. Suppose the random variable of interest is the total number of color televisions in two households selected at random. Compute the expected value and standard deviation of this new random variable.
Exercises

E1. A large class was assigned a difficult homework problem. Display 6.19 shows the scores the students received and the proportion of students who received each score.

<table>
<thead>
<tr>
<th>Score</th>
<th>Proportion of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.45</td>
</tr>
<tr>
<td>1</td>
<td>0.09</td>
</tr>
<tr>
<td>2</td>
<td>0.0</td>
</tr>
<tr>
<td>3</td>
<td>0.04</td>
</tr>
<tr>
<td>4</td>
<td>0.15</td>
</tr>
<tr>
<td>5</td>
<td>0.27</td>
</tr>
</tbody>
</table>

Display 6.19: Distribution of homework scores.

a. Compute the mean and standard deviation of the scores.
b. Describe how to use this list of random digits to take a sample of size 4 from these scores:
   30558  45957  36911  97199  08432
   (Note: The digits are grouped in batches of five to make them easier to read. They’re to be interpreted as a single string of digits.)

E2. Display 6.20 gives ages of students enrolled at the University of Texas, San Antonio.

<table>
<thead>
<tr>
<th>Age</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>17–22</td>
<td>44.2</td>
</tr>
<tr>
<td>23–29</td>
<td>31.4</td>
</tr>
<tr>
<td>30+</td>
<td>24.4</td>
</tr>
</tbody>
</table>


a. Make a reasonable estimate of the center of each age group, and compute the mean and standard deviation of the ages of these students.
b. Describe how to use this list of random digits to take a sample of six ages:
   30558  45957  36911  97199  08432

E3. Refer to Display 6.1 on page 359. If your boss provides two parking spaces per single-family house, how many vehicles do you estimate will have to be parked on the street in the new neighborhood of 500 single-family houses? What if your boss provides three spaces per house?

E4. Refer to Displays 6.1 and 6.3. Your boss suggests providing two parking spaces for each of the 500 single-family houses and, to keep things fair, four spaces for each of the 500 duplexes. If this plan is followed, in which neighborhood will a larger proportion of vehicles have to be parked on the street?

E5. About 10% of high school girls reported in a survey that they rarely or never wear seat belts in motor vehicles. Suppose you randomly sample three high school girls. Construct the probability distribution of the random variable, \( X \), defined as the number of girls in your sample who say they rarely or never wear seat belts. Begin by listing the outcomes as in the lung cancer example on page 362. [Source: www.nhtsa.dot.gov.]

E6. As stated in the example on page 362, 90% of lung cancer cases are caused by smoking. Suppose three lung cancer patients are randomly selected from the large population of people with that disease. Construct the probability distribution of \( X \), the number of patients with cancer caused by smoking.

E7. The report cited in E5 says that 18% of high school boys report that they rarely or never wear seat belts. Suppose you randomly sample two high school boys. Construct the probability distribution of the random variable, \( X \), defined as the number of sampled boys who say they rarely or never wear seat belts.

E8. According to a recent government report, 73% of drivers now use seat belts regularly. Suppose a police officer at a road check randomly stops three cars to check for seat belt usage. Construct the probability distribution of \( X \), the number of drivers using seat belts. [Source: usgovinfo.about.com.]
E9. Refer to the burglary insurance example on page 369.
   a. If you charge an urban householder $209.50 per year for insurance, what is the largest profit you could earn on that one policy for the year? What is the largest possible loss for the year?
   b. If you want to earn an expected yearly profit of $5000 per 1000 urban customers, how much should you charge per customer?
   c. Factors besides the location of the household affect the probability of a burglary. What other factors might an insurance company take into account?

E10. The passenger vehicle with the highest theft loss is the Cadillac Escalade EXT, with 20.2 claims for theft per 1000 insured vehicles per year and an average payment of $14,939 per claim. How much would you charge an owner of a Cadillac Escalade EXT for theft insurance per year if you simply wanted to expect to break even? [Source: Highway Loss Data Institute news release, October 19, 2004, www.hwysafety.org.]

E11. The boss of your construction company isn’t satisfied with the simulation in Display 6.3 on page 360 and asks you to construct the exact probability distribution for selecting two households at random and counting the total number of motor vehicles.
   a. Use the rules of probability from Chapter 5 and the information in Display 6.1 on page 359 to construct this distribution exactly.
   b. How does your exact distribution compare to the simulated distribution in Display 6.3?
   c. Find the mean value of this exact distribution and compare it to the expected value found in P17.

E12. Display 6.18 on page 378 provides information on the number of color televisions in U.S. households. Suppose a duplex is occupied by two randomly selected households and you are interested in the random variable X, defined as the total number of color televisions in a duplex.
   a. Construct the exact probability distribution for X.
   b. Find the expected value of X.
   c. What assumptions are necessary for your answer in part a?

E13. For each million tickets sold, the original New York Lottery awarded one $50,000 prize, nine $5,000 prizes, ninety $500 prizes, and nine hundred $50 prizes.
   a. What was the expected value of a ticket?
   b. The tickets sold for $0.50 each. How much could the state of New York expect to earn for every million tickets sold?
   c. What percentage of the income from the lottery was returned in prizes?

E14. A scratch-off card at a fast-food restaurant chain gave the information contained in Display 6.21. All cards were potential winners. In order to actually win the prize, the player had to scratch off a winning path without making a false step. The probability of doing this was given as \( \frac{1}{5} \) for the food item and \( \frac{1}{10} \) for any nonfood item. So, for example, if you got the card with the prize worth $1 million, you still would have to scratch off a path without making a mistake. The probability of doing that and actually winning the $1 million was \( \frac{1}{10} \).

<table>
<thead>
<tr>
<th>Value of Prize (in dollars)</th>
<th>Number of Game Cards</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,000,000</td>
<td>1</td>
</tr>
<tr>
<td>500,000</td>
<td>2</td>
</tr>
<tr>
<td>(house) 200,000</td>
<td>1</td>
</tr>
<tr>
<td>100,000</td>
<td>10</td>
</tr>
<tr>
<td>(car) 16,619</td>
<td>75</td>
</tr>
<tr>
<td>10,000</td>
<td>50</td>
</tr>
<tr>
<td>(cruise) 2,900</td>
<td>150</td>
</tr>
<tr>
<td>1,000</td>
<td>200</td>
</tr>
<tr>
<td>100</td>
<td>1,000</td>
</tr>
<tr>
<td>10</td>
<td>10,000</td>
</tr>
<tr>
<td>(food item) 0.73</td>
<td>196,000,000</td>
</tr>
</tbody>
</table>

Display 6.21 Values of prizes, along with frequencies.
a. What is the total value of all the potential prizes?
b. What total amount did the restaurant expect to pay out?
c. What is the expected value of a card?

E15. You are trying to decide whether to buy Brand A or Brand B as your new dishwasher. Each brand is expected to last about 10 years. Brand A has a price of $950 and comes with an unlimited number of repairs at $150 each. Brand B has a higher price, $1200, but comes with an unlimited number of free repairs. Which dishwasher you buy depends on the number of repairs you expect Brand A to require. You investigate and find out the information about Brand A in Display 6.22.

<table>
<thead>
<tr>
<th>Number of Repairs</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.40</td>
</tr>
<tr>
<td>1</td>
<td>0.30</td>
</tr>
<tr>
<td>2</td>
<td>0.15</td>
</tr>
<tr>
<td>3</td>
<td>0.10</td>
</tr>
<tr>
<td>4</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Display 6.22 Probability of number of repairs.

a. Find the expected cost of each brand, including original price and repairs.
b. What would be the advantage of buying Brand A? Brand B?

E16. Display 6.23 gives information about the four most commonly used local service plans for the New York/New Jersey calling area.

<table>
<thead>
<tr>
<th>Service</th>
<th>Monthly Charge</th>
<th>Monthly Minutes</th>
<th>Each Additional Minute</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT&amp;T</td>
<td>$49.99</td>
<td>600</td>
<td>$0.45</td>
</tr>
<tr>
<td>Verizon</td>
<td>$49.00</td>
<td>500</td>
<td>$0.45</td>
</tr>
<tr>
<td>Voicestream</td>
<td>$39.99</td>
<td>500</td>
<td>$0.25</td>
</tr>
<tr>
<td>Sprint PCS</td>
<td>$49.99</td>
<td>500</td>
<td>$0.30</td>
</tr>
</tbody>
</table>


If you talk for 500 minutes or less each month, Voicestream will be the cheapest service. But some months you have a lot to talk about, and you estimate that the number of minutes you talk each month follows the probability distribution in Display 6.24.

<table>
<thead>
<tr>
<th>Number of Minutes</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>500 or less</td>
<td>0.4</td>
</tr>
<tr>
<td>550</td>
<td>0.3</td>
</tr>
<tr>
<td>600</td>
<td>0.2</td>
</tr>
<tr>
<td>700</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Display 6.24 Probability of number of minutes.

Rank the service plans according to how much you would expect them to cost per month, knowing that your only calls will be within the New York/New Jersey calling area.

E17. Suppose you roll two tetrahedral dice, each with faces numbered 1, 2, 3, and 4.

a. Make a probability distribution table for the sum of the numbers on the two dice. What is the probability that the sum is 3?
b. Make a probability distribution table for the absolute value of the difference of the numbers on the two dice. What is the probability that the absolute value of the difference is 3?
c. What is the expected value of the probability distribution in part b?

E18. Construct the probability distribution table for the situation of flipping a coin five times and counting the number of heads. Compute the mean and standard deviation of the distribution, and construct a probability distribution graph.

E19. Many mechanical and electronic systems are built to include backups to major components (much like the spare tire in a car). If the main component fails, the backup kicks in. Often, however, the backup component is for emergency use only and is not built to the same specifications as the main component. Suppose a main pump in a city water system works without failure for 1 month with probability 0.1, 2 months with probability 0.3, and 3 months with probability 0.6. The pump’s backup works without failure for 1 month with probability...
0.2 and for 2 months with probability 0.8. The main pump is used alone until it fails; then the backup pump kicks in and is used alone until it fails. You can assume that the main pump and its backup operate independently of each other.

a. Find the probability distribution of the total working time of the main pump followed by its backup.
b. Find the expected value and standard deviation of the total working time of the main pump and its backup.
c. Show that the expected total time in part b is the sum of the expected working times of the main pump and of its backup.

E20. You pay $1 to play Game A, which generates a payoff of $0, $1, $2, or $3 with respective probabilities 0.4, 0.3, 0.2, and 0.1. You also pay $2 to play Game B, which generates a payoff of $0, $2, or $4 with respective probabilities 0.7, 0.2, and 0.1. The games are operated independently of each other.

a. Construct the probability distribution of your total gain from playing the two games. (Your gain is your winnings minus the cost of playing the game.)
b. Find the expected value and standard deviation of your total gain.
c. Show that your expected total gain in part b is the sum of your expected gains from each of the two games.

d. A regular die is rolled, and then a tetrahedral die is rolled. The difference is calculated.

a. Construct the probability distribution of the difference of the number on the regular die and the number on the tetrahedral die.
b. Verify that the mean of the probability distribution is \( \mu_1 - \mu_2 \) and the variance is \( \sigma_1^2 + \sigma_2^2 \), where \( \mu_1 \) and \( \sigma_1^2 \) are, respectively, the mean and variance of the distribution of the roll of a single six-sided die and \( \mu_2 \) and \( \sigma_2^2 \) are, respectively, the mean and variance of the distribution of the roll of a single four-sided die.

E22. A single die is rolled and the result used twice, once as the “first number” and again as the “second number.”

a. Construct the probability distribution of the difference of the first number and the second number.
b. Find the mean and variance of this probability distribution using the standard formulas

\[
\mu_X = \sum x_ip_i
\]

\[
\sigma_X^2 = \sum (x_i - \mu_X)^2p_i
\]

c. Is the mean of the differences equal to \( \mu_1 - \mu_2 \)?
d. Is the variance equal to \( \sigma_1^2 + \sigma_2^2 \)?

### 6.2 The Binomial Distribution

In Section 6.1, you used both theory and simulation to construct probability distributions of random variables. You might have found this to be a lot of work, even for relatively simple situations, and might have wondered if there was a simpler way to do it. Well, good news is coming. For some of the most commonly used random variables, mathematical formulas will give you the distribution directly. These types of situations have a common structure that makes a mathematical generalization possible. How can you find these magic formulas? In this section, Jack and Jill come to the rescue and show you how it’s done.
Binomial Probabilities

Many random variables seen in practice amount to counting the number of successes in \( n \) independent trials, such as

- the number of doubles in four rolls of a pair of dice (on each roll either you get doubles or you don't),
- the number of patients with type A blood in a random sample of ten patients (a person either has type A blood or doesn't)
- the number of defective items in a sample of 20 items (either an item is defective or it isn't)

These situations are called binomial because each trial has two possible outcomes.

In Chapter 5, Jack and Jill were asking questions about a binomial situation when they wanted to know the answer to a question like this: “If we ask four people which is the tap water and none of them can tell bottled water from tap water, what is the probability that two of the four will guess correctly?”

Jack: That's exactly what we were doing—looking at \( n \) people and counting the number who correctly identified the tap water.

Jill: Do they mean there is an easier way than listing all those outcomes?

Jack: It sounds like it. But I bet we can figure it out for ourselves now that we know there is a formula. We presented each of four people with bottled and tap water and asked them to identify the tap water. We got a list of 16 possible outcomes, all equally likely if none of the four can tell the difference and are merely guessing. Let's call the outcomes \( C \) for correctly identifying the tap water and \( I \) for incorrectly identifying it.

Jill: Yeah, look. I’ve organized them in this table (Display 6.25). Let's try to figure out the probability of getting exactly two people who select the tap water.

<table>
<thead>
<tr>
<th>Number Who Correctly Identify Tap Water</th>
<th>Outcomes, ( C ) (correct), ( I ) (incorrect)</th>
<th>Number of Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>IIII</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>IICC  IICI  ICII  CIII</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>IICC  ICIC  ICIC  CICI  CICI  CICI</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>ICCC  CICC  CCIC  CCCI</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>CCCC</td>
<td>1</td>
</tr>
</tbody>
</table>

Display 6.25  Jill’s table.

Jack: Using the Multiplication Rule for Independent Events, we know that the probability of getting a particular outcome when two people guess correctly, say, \( CIC \) is \( \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \) or \( \frac{1}{16} \). There are six outcomes where two people guess correctly. So the probability of getting two people who correctly select tap water is \( \frac{6}{16} \). But I wish we could find the numbers in the last column without listing all the possible outcomes!
Jill: I've seen those numbers before—in the fifth row of Pascal's triangle.

Jack: That's it! The number of ways we can choose exactly two people out of four people to identify the tap water is

$$\binom{4}{2} = \frac{4!}{2!2!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 2 \cdot 1} = 6$$

DISCUSSION

Binomial Probabilities

D13. Explain why each of the 16 outcomes in Display 6.25 has probability $\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}$.

Example: College Graduates

The proportion of adults age 25 and older in the United States with at least a bachelor’s degree is 0.27. [Source: U.S. Census Bureau, Statistical Abstract of the United States, 2006, Table 216.] Suppose you pick seven adults at random. What is the probability that exactly three will have a bachelor’s degree or higher?

Solution

Jack: Now what do we do? They’ve changed the problem so that the outcomes aren’t equally likely. That is so like them.

Jill: There is nothing about our method that requires equally likely outcomes—all that matters is independence and that the probability stays the same. For example, one outcome with three grads is

grad not grad grad not not not

By the Multiplication Rule, the probability of this particular outcome is

$$(0.27)(0.73)(0.27)(0.73)(0.73)(0.73)(0.73) = (0.27)^3(0.73)^4$$

Another outcome with three grads is

not grad grad not not not grad

The probability of this particular outcome is

$$(0.73)(0.27)(0.27)(0.73)(0.73)(0.73)(0.27) = (0.27)^3(0.73)^4$$

which is exactly the same.

Jack: That’s because the probabilities of the outcomes with exactly three grads all have the same factors but in a different order.

Jill: And there are 35 of them because

$$\binom{7}{3} = \frac{7!}{3!4!} = 35$$

It’s a good thing we didn’t have to list all of them!
Jack: So the probability of getting exactly three college grads is 
(number of ways to get 3 grads) \cdot (probability of each way) 
or
\[ \binom{7}{3}(0.27)^3(0.73)^4 \approx 0.196 \]

Jill: Yeah! Now we can do any problem they throw at us.

A summary of Jack and Jill’s method is given next.

### The Binomial Probability Distribution

Suppose you have a series of trials that satisfy these conditions:

- **B:** They are binomial—each trial must have one of two different outcomes, one called a “success” and the other a “failure.”
- **I:** Each trial is independent of the others; that is, the probability of a success doesn’t change depending on what has happened before.
- **N:** There is a fixed number, \( n \), of trials.
- **S:** The probability, \( p \), of a success is the same on each trial, with \( 0 < p < 1 \).

Then the distribution of the random variable \( X \) that counts the number of successes is called a **binomial distribution**. Further, the probability that you get exactly \( X = k \) successes is

\[
P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}
\]

where \( \binom{n}{k} \) is the binomial coefficient, calculated as

\[
\binom{n}{k} = \frac{n!}{k!(n-k)!}
\]

where \( n! = n(n-1)(n-2) \cdots 3 \cdot 2 \cdot 1 \).

You can use a calculator to calculate binomial probabilities quickly. For example, the probability in the previous example—that exactly three of seven randomly chosen adults will have at least a bachelor’s degree—is shown here. [See Calculator Note 6B to learn more about this calculator function.]

\[
\text{binompdf}(7, 0.27, 3) = 0.1956369022
\]

You can also calculate cumulative probabilities using a calculator. For example, the probability that fewer than three (that is, zero, one, or two) of seven
adults will have a bachelor’s degree is shown here. [See Calculator Note 6C to learn how to perform this operation.]

You might have noticed that the trials in Jack and Jill’s example aren’t really independent. The first adult they selected who was age 25 or older has probability 0.27 of being a college graduate. If that person is a college graduate, the probability that the next person selected is a college graduate is a bit less. However, the change in probability is so small that Jack and Jill can safely ignore it. If there are 150,000,000 adults age 25 and older in the United States, then there would be 40,500,000 college graduates. The probability that the first adult selected is a college graduate is \( \frac{40,500,000}{150,000,000} \), or 0.27. If that person turns out to be a college graduate, the probability that the second adult selected is a college graduate is \( \frac{40,499,999}{149,999,999} \), or 0.2699999951, which is very close to 0.27.

You can treat your random sample as a binomial situation as long as the sample size, \( n \), is small compared to the population size, \( N \). A simple guideline that works well in practice is that \( n \) should be less than 10% of the size of the population, or \( n < 0.10N \).

The Binomial Probability Distribution

D14. Make a probability distribution table for the number of college graduates in a random sample of seven adults if 27% of adults are college graduates.

D15. Make a probability distribution table for the number of people who aren't college graduates. Make a histogram of this distribution and of the distribution in D14, and compare the two.

D16. Show why the population must be “large” in order to use the binomial probability formula by pretending that the total adult population of the United States is 12, of which 4 are “chocoholics.” Compute the exact probability that in a random sample of 7 adults, none will be chocoholics. Compare your results to those you get using the binomial formula.

D17. From a class of 20 students, half of whom are seniors, the teacher selects 10 students at random to check for completion of today’s paper. If \( X \) denotes the number of papers belonging to seniors among those checked, will \( X \) have a binomial distribution? Explain why or why not.

In Activity 6.2a, you’ll conduct the tap water vs. bottled water experiment yourself. This activity will give you practice in statistical decision-making as you decide how many subjects will have to correctly select the tap water before you are convinced that people can tell the difference.

DISCUSSION

In a famous episode of I Love Lucy, Lucille Ball takes her passion for chocolate to comic extremes.
ACTIVITY 6.2a Can People Identify the Tap Water?

**What you'll need:** small paper or plastic cups, a container of tap water, a container of bottled water, about 20 volunteer subjects

1. Design a study to see whether people can identify which of two cups of water contains the tap water. Refer to the principles of good experimental design in Chapter 4.
2. Suppose none of your 20 subjects actually can tell the difference, so their choice is equivalent to selecting one of the cups at random. Construct a probability distribution of the number who choose the tap water.
3. If only ten people correctly select the tap water, you have no reason to conclude that people actually can tell the difference. Why?
4. Decide how many people will have to make the correct selection before you are reasonably convinced that people can tell the difference.
5. Conduct your study and write a conclusion.

### Shape, Center, and Spread of a Binomial Distribution

The fact that the binomial distribution is represented by such a simple and elegant formula might lead you to believe that there is a simple formula for the expected value and standard deviation as well. Jack and Jill to the rescue again! For Jack and Jill’s tap water tasting experiment, the distribution of $X$, the number choosing tap water out of four tasters, is shown in Display 6.26.

<table>
<thead>
<tr>
<th>$k$</th>
<th>$P(X = k)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1/16</td>
</tr>
<tr>
<td>1</td>
<td>4/16</td>
</tr>
<tr>
<td>2</td>
<td>6/16</td>
</tr>
<tr>
<td>3</td>
<td>4/16</td>
</tr>
<tr>
<td>4</td>
<td>1/16</td>
</tr>
</tbody>
</table>

*Display 6.26  Distribution of number of correct guesses.*

The expected value for this distribution is

$$E(X) = 0(1/16) + 1(4/16) + 2(6/16) + 3(4/16) + 4(1/16) = 2$$

Notice that $E(X)$ can be written as $4 \cdot 1/2$, which is $np$. In other words, you would expect about half of the four to guess correctly given two equally likely choices.

Is this pattern just coincidence, or is it a general rule? Let’s try one more example to see if we have any hope of the latter. The discussion of college...
graduates on page 384 stated that about 27% of U.S. adults have at least a bachelor’s degree. If three adults are randomly sampled in an opinion poll, the distribution of \( X \), the number with at least a bachelor’s degree, will be as shown in Display 6.27.

<table>
<thead>
<tr>
<th>( k )</th>
<th>( P(X = k) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>((0.73)^3 = 0.39)</td>
</tr>
<tr>
<td>1</td>
<td>(3(0.27)(0.73)^2 \approx 0.43)</td>
</tr>
<tr>
<td>2</td>
<td>(3(0.27)^2(0.73) \approx 0.16)</td>
</tr>
<tr>
<td>3</td>
<td>((0.27)^3 = 0.020)</td>
</tr>
</tbody>
</table>

Display 6.27 Distribution of number of college graduates.

The expected value of this distribution is

\[
E(X) = 0(0.39) + 1(0.43) + 2(0.16) + 3(0.02) = 0.81 = 3(0.27) = np
\]

A general rule does seem to be a possibility, and that, indeed, is the case. As a bonus, there is a general rule for the standard deviation that turns out to be just about as simple as the one for the expected value.

The Characteristics of a Binomial Distribution

For a random variable \( X \) having a binomial distribution with \( n \) trials and probability of success \( p \), the mean (expected value) and standard deviation for the distribution are given by

\[
E(X) = \mu_X = np \quad \text{and} \quad \sigma_X = \sqrt{np(1 - p)}
\]

The shape of the distribution becomes more normal as \( n \) increases.

Example: College Graduates

About 27% of U.S. adults have at least a bachelor’s degree. If you select 100 adults at random from all adults in the United States, how many do you expect to have a bachelor’s degree and with what standard deviation?

Solution

You expect to get \( E(X) = 100(0.27) \), or about 27 college graduates, with a standard deviation of \( \sigma_X = \sqrt{100(0.27)(1 - 0.27)} \), or 4.44.

Plots of binomial distributions for various values of \( n \) and \( p \) are shown in Display 6.28.
6.2 The Binomial Distribution

You can use a calculator to quickly graph and compare binomial distributions. [See Calculator Note 6D to learn how.] The binomial distribution for \( n = 5 \) and \( p = 0.7 \) is shown here. In D18 you’ll explore how binomial distributions change for increased values of \( p \) or \( n \).

Shape, Center, and Spread of a Binomial Distribution

D18. How do the shape, center, and spread of a binomial distribution change as \( p \) increases for fixed values of \( n \)? How do they change as \( n \) increases for fixed values of \( p \)?

D19. Ten CD players of a discontinued model are selling for $100 each, with a “double your money back” guarantee if they fail in the first month of use. Suppose the probability of such failure is 0.08. What is the expected net gain for the retailer after all ten CD players are sold? Ignore the original cost of the CD players to the retailer.

Summary 6.2: The Binomial Distribution

The random variable \( X \) is said to have a binomial distribution if \( X \) represents the number of “successes” in \( n \) independent trials, where the probability of a success is \( p \) on each trial. For example, \( X \) might represent the number of successes in a random sample of size \( n \) from a large population, with probability of success \( p \) on each selection.
A binomial distribution has these important characteristics:

- The probability of getting exactly $X = k$ successes is given by the formula
  
  $$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

- The expected value or mean is $E(X) = \mu_X = np$.
- The standard deviation is $\sigma_X = \sqrt{np(1 - p)}$.

### Practice

#### Binomial Probabilities

P19. Suppose Jack and Jill ask six people who can’t tell the difference between tap water and bottled water to identify the tap water. Use their method to make a probability distribution table for six people. Then make a graph of the distribution.

P20. Suppose you flip a coin eight times. What is the probability that you’ll get exactly 3 heads? Exactly 25% heads? At least 7 heads?

P21. Suppose you roll a balanced die seven times. What is the probability that you will get an even number exactly two times? More than half the time?

### The Binomial Probability Distribution

P22. About 8.8% of people ages 14–24 are “dropouts,” persons who are not in regular school and who have not completed the 12th grade or received a general equivalency degree. Suppose you pick five people at random from this age group. [Source: U.S. Census Bureau, Statistical Abstract of the United States, 2006, Table 255.]

a. What is the probability that none of the five are dropouts?

b. What is the probability that at least one is a dropout?

c. Make a probability distribution table for this situation.

P23. Describe how to use simulation to construct an approximate binomial distribution for the situation in P22. Conduct two trials of your simulation, using these random digits:

89254 99538 18315 45716 36270 79665
49830 06226 88863 02322 36630 07176

P24. According to a recent government report, 73% of drivers now use seat belts regularly. Suppose a police officer at a road check randomly stops four cars to check for seat belt usage. Find the probability distribution of $X$, the number of drivers using seat belts. [Source: National Highway Traffic Safety Administration.]

#### Shape, Center, and Spread of a Binomial Distribution

P25. The median annual household income for U.S. households is about $44,400. [Source: U.S. Census Bureau, www.census.gov.]

a. Among five randomly selected U.S. households, find the probability that four or more have incomes exceeding $44,400 per year.

b. Consider a random sample of 16 U.S. households.

   i. What is the expected number of households with annual incomes under $44,400?

   ii. What is the standard deviation of the number of households with annual incomes under $44,400?

   iii. What is the probability of getting at least 10 out of the 16 households with annual incomes under $44,400?

c. In a sample of 16 U.S. households, suppose none had annual incomes under $44,400. What might you suspect about this sample?
P26. You neglected to study for a true–false quiz on the government of Botswana, so you will have to guess on all the questions. You need to have at least 60% of the answers correct in order to pass. Would you rather have a 5-question quiz or a 20-question quiz? Explain your reasoning, referring to Display 6.28.

### Exercises

E23. If you roll a pair of dice five times, find the probability of each outcome.
   a. You get doubles exactly once.
   b. You get exactly three sums of 7.
   c. You get at least one sum of 7.
   d. You get at most one sum of 7.

E24. Suppose you select five numbers at random from 10 through 99, with repeats allowed. Find the probability of each outcome.
   a. Exactly three of the numbers are even.
   b. Exactly one of the numbers has digits that sum to a number greater than or equal to 9.
   c. At least one of the numbers has digits that sum to a number greater than or equal to 9.

E25. According to a recent Census Bureau report, 37 million Americans, or 12.7% of the population, live below the poverty level. Suppose these figures hold true for the region in which you live. You plan to randomly sample 25 Americans from your region. [Source: Current Population Survey, 2005 Annual Social and Economic Supplement, www.census.gov.]
   a. What is the probability that your sample will include at least two people who do not have health insurance?
   b. What are the expected value and standard deviation of the number of people in your sample without health insurance?

E26. According to the U.S. Census Bureau, about 16% of residents have no health insurance. You are to randomly sample 20 residents for a survey on health insurance coverage. [Source: U.S. Census Bureau, Current Population Survey, March 2004.]
   a. What is the probability that your sample will include at least three people who do not have health insurance?
   b. What are the expected value and standard deviation of the number of people in your sample without health insurance?

E27. You buy 15 lottery tickets for $1 each. With each ticket, you have a 0.06 chance of winning $10. Taking into account the cost of the tickets,
   a. what is your expected gain (or loss) on this purchase?
   b. what is the probability that you will gain $10 or more?
   c. what is the standard deviation of your gain?

E28. An oil exploration firm is to drill ten wells, each in a different location. Each well has a probability of 0.1 of producing oil. It will cost the firm $60,000 to drill each well. A successful well will bring in oil worth $1 million. Taking into account the cost of drilling,
   a. what is the firm's expected gain from the ten wells?
   b. what is the standard deviation of the firm's gain for the ten wells?
   c. what is the probability that the firm will lose money on the ten wells?
   d. what is the probability that the firm will gain $1.5 million or more from the ten wells?

E29. A home alarm system has one detector for each of the $n$ zones of the house. Suppose the probability is 0.7 that the detector sounds an alarm when an intruder passes through its zone and that this probability is the
same for each detector. The alarms operate independently. An intruder enters the house and passes through all the zones.

a. What is the probability that at least one alarm sounds if \( n = 3 \)?

b. What is the probability that at least one alarm sounds if \( n = 6 \)? Is the probability from part a doubled?

c. How large must \( n \) be in order for the probability of at least one alarm sounding to be about 0.99?

E30. Complex electronic systems for which failure could be catastrophic (such as airplane electronic systems) are constructed so that key components have a number of parallel backup components that automatically take over in case of failure. Suppose that at a certain key juncture there are \( n \) electronic switches (the main one and \( n - 1 \) backups) that operate independently of one another under standard conditions. The probability that any one switch works properly is 0.92.

a. What is the probability that at least one switch works properly if \( n = 2 \)?

b. What is the probability that at least one switch works properly if \( n = 3 \)?

c. An engineer wants the probability that at least one switch works properly to be \( 1 - 10^{-5} \), or 0.99999. How many backup switches should she place in the system?

E31. A potential buyer will sample DVDs from a large lot of new DVDs. If she finds at least one defective DVD, she'll reject the entire lot. In each case, find the sample size \( n \) for which the probability of detecting at least one defective DVD is 0.50.

a. Ten percent of the DVDs are defective.

b. Four percent of the DVDs are defective.

e. Show that the sum of the probabilities when \( n = 2 \) is \( p^2 + 2pq + q^2 \). Factor this sum to show that it is equal to 1.

b. Find the sum of the probabilities when \( n = 3 \). Factor this sum to show that it is equal to 1.

c. Show that the terms of the expansion of \((p + q)^n\) are the probabilities for \( k = 0, 1, 2, \ldots, n \) in the binomial distribution with \( n \) trials and probability of success \( p \).
ACTIVITY 6.3a

Waiting for Type A Blood

What you'll need: a device for generating random outcomes

In the general population, about 40% of people have type A blood. [Source: American Red Cross, September 2006, www.givelife2.org.] A technician in a blood bank needs type A blood today and wants to know about how many randomly selected blood donations he might have to process to find the first donation that is type A.

1. Describe how to use random digits to simulate the blood type of the first donation he checks.
2. Simulate the outcome for the first donation he checks. Was it type A blood? In other words, was his waiting time to success just one trial?
3. If he was successful on the first donation, then stop this run of the simulation. If he did not have success on the first trial, then continue generating outcomes until he gets his first success. Record the number of the trial on which the first success occurred.
4. Repeat steps 2 and 3 at least ten times, each time recording the number of trials needed.
5. Combine your results with others from the class and construct a plot to represent the simulated distribution of the number of donations that must be checked. Describe the shape of this distribution, and find its mean.
6. What is your estimated probability that
   a. the first donation with type A blood is the second one he checks?
   b. he will have to check five or fewer donations to find the first that is type A?
   c. he has to check at least two donations to get one that is type A?
   d. he has to check at most four donations to get one that is type A?
7. On which trial is he most likely to find the first donation that is type A?
The probability distribution you constructed in the activity is called a geometric distribution. Display 6.29 shows geometric distributions for $p$ equal to 0.1, 0.3, 0.5, and 0.8, where $p$ is the probability of “success” on each trial. For example, when $p = 0.3$, the probability of getting a success on the first trial is 0.3, as the height of the first bar indicates. The probability of getting the first success on the second trial is 0.21, the height of the second bar. In this case, you must fail on the first try and then succeed on the second, and this probability is

$$P(\text{fail on first try}) \cdot P(\text{succeed on second try}) = (0.7)(0.3) = 0.21$$

Note that in each graph, each bar is the same fixed proportion of the height of the bar on its left.

![Geometric distributions for different probabilities of a success.](image)

[See Calculator Note 6E to learn how to graph geometric distributions.]

**DISCUSSION**

**Waiting-Time Situations**

D20. Refer to the graphs in Display 6.29.

a. What is the shortest possible waiting time? What is the longest? What is the most likely waiting time? Why is this the case?

b. What are the exact heights of the first and second bars in the graph for $p = 0.8$? Explain.

c. How are the shapes of the four distributions in Display 6.29 similar? How are they different?

d. What real-life situation could the graph for $p = 0.5$ represent?
The Formula for a Geometric Distribution

As you might already have discovered, you can derive a formula for the probability distribution of a geometric (waiting-time) random variable quite easily.

Jack: Yep, we can do it. This is easier than our problem from Section 6.2 about the binomial distribution.

Jill: It sure is. Let's do it for the blood bank example, where about 40% of people have type A blood. There the probability that the first donation checked is type A is 0.4.

Jack: Right. The probability that the waiting time, \( X \), is 1 is \( P(X = 1) = 0.4 \).

Jill: Now for \( P(X = 2) \). For the second donation to be the first that is type A . . .

Jack: What kind of nonsense is that? “For the second donation to be the first . . . ?”

Jill: Sorry, let me say it with more words. Suppose the first donation checked isn't type A and the second donation is. Then we have our first success with the second donation.

Jack: That's better. The probability of this sequence of outcomes is \( P(\text{first donation isn't type A and second is type A}) = (0.6)(0.4) \), or 0.24.

Jill: But you get to multiply like that only if the events are independent. What if a whole family had donated blood? Then their blood types might not be independent.

Jack: Yeah, we will have to be careful about things like that. The probability can't change depending on who else has come in.

Jill: If we can assume independence, the probability that the third donation checked will be the first that is type A is \( P(X = 3) = (0.6)(0.6)(0.4) \).

Jack: Because we have to have two “failures” and then our first success.

Jill: This could go on forever. So we better get started and make a table.

<table>
<thead>
<tr>
<th>Number of Trials to Get First Success</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.4</td>
</tr>
<tr>
<td>2</td>
<td>((0.6)(0.4) = 0.24)</td>
</tr>
<tr>
<td>3</td>
<td>((0.6)(0.6)(0.4) = 0.144)</td>
</tr>
<tr>
<td>4</td>
<td>((0.6)(0.6)(0.6)(0.4) = 0.0864)</td>
</tr>
<tr>
<td>5</td>
<td>((0.6)(0.6)(0.6)(0.6)(0.4) = 0.05184)</td>
</tr>
</tbody>
</table>

Display 6.30  Jill’s table.

Jack: All this work is making me thirsty. Five rows is enough. Let's go up the hill and get some water.
Display 6.31 summarizes Jack and Jill’s reasoning, this time for a general geometric distribution with probability of success \( p \) and probability of failure \( q = 1 - p \).

<table>
<thead>
<tr>
<th>Number of Trials to Get First Success</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( p )</td>
</tr>
<tr>
<td>2</td>
<td>( q \cdot p )</td>
</tr>
<tr>
<td>3</td>
<td>( q^2 \cdot p )</td>
</tr>
<tr>
<td>4</td>
<td>( q^3 \cdot p )</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>( k )</td>
<td>( q^{k-1} \cdot p )</td>
</tr>
</tbody>
</table>

**Display 6.31**  Table for the general geometric distribution.

### The Geometric Probability Distribution

Suppose you have a series of trials that satisfy these conditions:

- They are binomial—that is, each trial must have one of two different outcomes, one called a “success” and the other a “failure.”
- Each trial is independent of the others; that is, the probability of a success doesn’t change depending on what has happened before.
- The trials continue until the first success.
- The probability, \( p \), of a success is the same on each trial, \( 0 < p < 1 \).

Then the distribution of the random variable \( X \) that counts the number of trials needed until the first “success” is called a **geometric distribution**. The probability that the first success occurs on the \( X = k \) th trial is

\[
P(X = k) = (1 - p)^{k-1} p
\]

for \( k = 1, 2, 3, \ldots \).

Both the binomial and geometric random variables start with a sequence of independent trials with two outcomes and constant probability of success. The binomial counts the number of successes in a fixed number of trials; the geometric counts the number of the trial on which the first success occurs. [See Calculator Notes 6F and 6G to learn how to calculate geometric probabilities and cumulative probabilities of a geometric distribution.]

### The Formula for a Geometric Distribution

D21. About 10% of the U.S. population has type B blood. [Source: American Red Cross, September 2006, www.givelife2.org.] Suppose our technician is checking donations that may be considered independent with respect to blood type.
6.3 The Geometric Distribution

a. What is the probability that the first donation that is type B is the third one checked?
b. What is the probability that he will have to check at most three donations to get the first that is type B? At least three donations?
c. Suppose that the technician is looking for a type B donation and has already tested three donations without success. What is the probability that he has to test at least three more donations to find one that is type B?
d. Part c refers to the memoryless property of the geometric distribution. Explain why that term is appropriate.

D22. In Activity 3.5a, Copper Flippers, you started by flipping 200 pennies. You then removed the ones that “died” (landed tails) and continued flipping with a decreasing number of “live” copper flippers. Make a plot with the number of the flip on the horizontal axis and the theoretical number of copper flippers left after that flip on the vertical axis. Compare this plot with your results from Activity 3.5a if you still have them.

D23. Give at least two ways to find the probability that it will take three or more rolls of a pair of dice to get doubles.

**Expected Value and Standard Deviation**

In Activity 6.3b, you will discover the formula for the mean (expected value) of a geometric distribution.

**ACTIVITY 6.3b**

**The Expected Value of a Geometric Distribution**

Divide your class into nine groups. Each group will be assigned a different value of $p$, for $p = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, \text{and} 0.9$.

1. Make a table that shows the probabilities of the values of $X$ from 1 through 30 for the geometric distribution with your group’s assigned value of $p$.
2. Use your table to compute an approximation of the expected value of this distribution.
3. Will your approximation in step 2 be a bit larger or a bit smaller than the theoretical expected value? Explain.
4. Using the results from all nine groups, make a scatterplot of the expected values plotted against the values of $p$. What is the shape of this plot?
5. Now make a scatterplot of the expected values plotted against the reciprocals of the values of $p$. State a hypothesis about the formula for the expected value of a geometric distribution.

Like the formulas for the binomial distribution, those for the expected value and standard deviation of the geometric distribution turn out to be surprisingly simple. Although using the mean as the measure of center and the standard deviation as the measure of spread seemed complicated in Chapter 2, it pays off now.
Characteristics of a Geometric Distribution

A random variable $X$ that has a geometric distribution with probability of success $p$ has an expected value (mean) and standard deviation of

$$\mu_X = \frac{1}{p} \quad \text{and} \quad \sigma_X = \frac{\sqrt{1-p}}{p}$$

The shape of the distribution is skewed right, with bars decreasing proportionally.

You can find a proof of the formula for the expected value in E42. The proof of the formula for the standard deviation is a bit more involved.

**Example: Mean and Standard Deviation for Blood Donations**

In the scenario described in Activity 6.3a, find the expected value and standard deviation of the number of blood donations the technician would have to check in order to get the first donation with type A blood.

**Solution**

The probability that a random blood donation is type A is 0.4. Thus, the expected number of blood donations that the technician would have to check in order to get the first one that is type A is

$$\mu_X = \frac{1}{p} = \frac{1}{0.4} = 2.5$$

with standard deviation

$$\sigma_X = \frac{\sqrt{1-p}}{p} = \frac{\sqrt{1-0.4}}{0.4} = 1.94$$

Note that the expected value is the expected number of donations that must be checked to get one that is type A. It is not the expected number checked before getting one that is type A. In this case, the expected number includes 1.5 donations that aren’t type A and one donation that is.

**Example: The Expected Number for the Second Success**

In the scenario described in Activity 6.3a, find the expected value of the number of blood donations the technician would have to check in order to get two that are type A blood.

**Solution**

The expected number of blood donations that must be checked to obtain two of type A is 2(2.5), or 5, because the technician must check, on average, 2.5 donations to get the first success. Then the process starts again to get the second success, again checking an average of 2.5 donations, or five in all.
6.3 The Geometric Distribution

Expected Value and Standard Deviation

D24. The phone line for a 24-hour ticket office is busy about 70% of the time.
   a. If you dial at random times throughout the week, what is the expected number of tries it will take you to get through? What is the standard deviation of the number of tries?
   b. What is the expected total number of times you will have to dial if you forget to ask a question after getting through the first time and have to call again?
   c. Suppose you have been trying for 3 days, so your friend starts dialing as well—at random times—while you continue trying. Who do you expect to get through first, you or your friend?

D25. If \( p \) doubles, how does the expected value of a geometric random variable change?

Summary 6.3: The Geometric Distribution

For a sequence of independent trials in which \( p \), the probability of “success,” stays the same for each trial, the random variable \( X \) that counts the number of trials until the first success occurs has a geometric distribution.

- The probabilities for this distribution are given by the formula
  \[
  P(X = k) = (1 - p)^{k-1}p
  \]
  for \( k = 1, 2, 3, \ldots \) and \( 0 < p < 1 \).
- The mean (expected value) of the distribution is
  \[
  E(X) = \mu_x = \frac{1}{p}
  \]
- The standard deviation is
  \[
  \sigma_x = \frac{\sqrt{1 - p}}{p}
  \]
- The expected number of trials until the \( n \)th success is \( n \cdot \frac{1}{p} \).

Practice

Waiting-Time Situations

P27. Suppose you are rolling a pair of dice and waiting for a sum of 7 to occur.
   a. What is the probability that you get a sum of 7 for the first time on your first roll? On your second roll?
   b. Using the graphs in Display 6.29 as a guide, sketch an approximate graph of the probability distribution of this situation.

The Formula for a Geometric Distribution

P28. About 85% of Americans over age 25 have graduated from high school. You are randomly sampling and interviewing adults one at a time for an opinion poll that applies only to high school graduates. [Source: U.S. Census Bureau, Statistical Abstract of the United States, 2006.]
   a. What is the probability that you get your first high school graduate on the third person you interview?
b. What is the probability that you get your first high school graduate sometime after the second person you interview?
c. Sketch an approximate distribution of the number of the interview on which you get the first high school graduate.

P29. Suppose 9% of the engines manufactured on a certain assembly line have at least one defect. Engines are randomly sampled from this line one at a time and tested. What is the probability that the first nondefective engine is found
a. on the third trial?
b. before the fourth trial?

d. What assumption are you making in computing these probabilities?

Expected Value and Standard Deviation

P30. About 70% of the time, the telephone lines coming into a concert ticket agency are all busy. Suppose you are calling this agency.

a. What is the probability that it takes you only one try to get through?
b. What is the probability that it takes you two tries?
c. What is the probability that it takes you four tries?
d. What assumption are you making in computing these probabilities?

d. What assumption are you making in computing these probabilities?

P31. The probability that a random blood donation is type B is 0.1.

a. What is the expected number of donations that must be checked to obtain the first that is type B?
b. What is the standard deviation of the number of donations that must be checked to obtain the first of type B?
c. What is the expected number of donations that must be checked to obtain two that are type B? To obtain three that are type B?

P32. About 85% of Americans over age 25 have graduated from high school. You are randomly sampling and interviewing adults one at a time for an opinion poll that applies only to high school graduates.

a. What is the expected number of interviews you have to conduct to get the first high school graduate?
b. What is the standard deviation of the number of interviews you have to conduct to get the first high school graduate?
c. What is the expected number of interviews you have to conduct to get the first ten high school graduates?
d. What is the standard deviation of the number of interviews you have to conduct to get the first ten high school graduates?

Exercises

E35. You are participating in a “question bee” in history class. You remain in the game until you give your first incorrect answer. The questions are all multiple choice, each with four possible answers exactly one of which is correct. Unfortunately, you have not studied for this bee, and you simply guess randomly on each question.

a. What is the probability that you are still in the bee after the first round?
b. What is the probability that you are still in the bee after the third round?
c. What is the expected number of rounds you will be in the bee?
d. If an entire class of 32 is simply guessing on each question, how many students
do you expect to be left after the third round?

E36. A certain type of insect has a 0.2 chance of dying on any one day and this probability remains relatively constant for a few days.

a. What is the probability that an insect of this type survives its first day?
b. What is the probability that an insect of this type survives its first three days?
c. What is the probability that an insect of this type dies on its fourth day?
d. How many days do you expect an insect of this type to survive?
e. If 200 insects of this type are being studied in a laboratory, how many do you expect to be alive after the third day?

E37. Suppose 12% of the engines manufactured on a certain assembly line have at least one defect. Engines are randomly sampled from this line one at a time and tested.

a. What is the expected number of engines that need to be tested to find the first engine without a defect? The third?
b. What is the standard deviation of the number of engines that need to be tested to find the first engine without a defect?
c. If it costs $100 to test one engine, what are the expected value and standard deviation of the cost of inspection up to and including the first engine without a defect?
d. Will the cost of inspection often exceed $200? Explain.

E38. An oil exploration firm is to drill wells at a particular site until it finds one that will produce oil. Each well has probability 0.1 of producing oil. It costs $60,000 to drill a well.

a. What is the expected number of wells to be drilled?
b. What is the standard deviation of the number of wells to be drilled?
c. What are the expected value and standard deviation of the cost of drilling to get the first successful well?
d. What is the probability that it will take at least five tries to get the first successful well? At least 15 tries?

E39. In a certain company, an employee must take an aptitude test before being promoted to the next job level, and the test cannot be taken until the employee has 6 months of experience in the company. Aptitude tests, however, can be designed to have an approximate probability, \( p \), that any one employee with 6 months or more experience will pass. What value of \( p \) should be the target of the test designer if the goal is to have 80% of the employees pass in no more than two tries? What assumption is necessary for your answer to be correct?

E40. An engineer is designing a machine to find tiny cracks in airplane wings that are difficult to detect. The specifications state that a section of wing known to have these cracks may be tested by the machine up to five times, but the probability of locating the cracks in five trials or fewer must be at least 0.99. In order to meet this specification, what must be the probability of locating cracks on any one trial?

E41. A geometric sequence is a sequence of the form

\[ a, ar, ar^2, ar^3, ar^4, \ldots \]

where \( a \) is the first term and \( r \) is the common ratio.

a. Explain why the probability distribution developed in this section is called the geometric distribution.
b. The sum of the probabilities in a geometric distribution form an infinite series:

\[ p + pq + pq^2 + pq^3 + pq^4 + \cdots \]

(See Display 6.31). Use the next formula, for the sum of an infinite geometric series, to prove that the sum of the probabilities of a geometric distribution is equal to 1 as long as \(|r| < 1\):

\[ a + ar + ar^2 + ar^3 + \cdots = \frac{a}{1 - r} \]
E42. In this exercise, you will use the last formula in E41 twice to prove that the expected value of the geometric distribution is \( \frac{1}{p} \).

a. Use the formula for the expected value of a probability distribution to write the formula for the mean of a geometric distribution with probability of success \( p \). This will be an infinite series.

b. Show that you can write the series from part a in this form:

\[
E(X) = p(1 + q + q^2 + q^3 + \cdots \\
+ q^2 + q^3 + \cdots \\
+ q^3 + \cdots \\
+ \cdots)
\]

c. Ignore \( p \) for now and sum each row using the last formula in E41.

d. Find the sum of all the rows, again using the last formula in E41.

---

**Chapter Summary**

In this chapter, you have learned that random variables are variables with a probability attached to each possible numerical outcome. Probability distributions, like data distributions, are characterized by their shape, center, and spread. You can use similar formulas for the mean and standard deviation.

The binomial distribution is a model for the situation in which you count the number of successes in a random sample of size \( n \) from a large population. A typical question is "If you perform 20 trials, what is the probability of getting exactly 6 successes?"

The geometric distribution is a model for the situation in which you count the number of trials needed to get your first success. A typical question is "What is the probability that it will take you exactly five trials to get the first success?"

---

**Review Exercises**

E43. Suppose you roll a dodecahedral (12-sided) die twice. Your summary statistic will be the sum of the two rolls.

a. What is the probability that the sum is 3 or less?

b. Compute the mean and standard deviation of the distribution of outcomes of a single roll of one die.

c. Compute the mean and standard deviation of the probability distribution of the sum of two rolls.

E44. Suppose you and your partner each roll two dice. Each of you computes the average of your two rolls. The summary statistic is the difference between your average and your partner’s average. Describe the sampling distribution of these differences.

E45. Two different pumping systems on levees have pumps numbered 1, 2, 3 and 4, configured as in Display 6.32. In System I, water will flow from A to B only if both pumps are working. In System II, water will flow from A to B if either pump is working. Assume that each pump has this lifelength distribution: 1 month with probability 0.1, 2 months with probability 0.3, and 3 months with probability 0.6. Lifelength refers to the length of time the pump will work continuously without repair. Assume that the pumps operate independently of each other. (You are interested only in whether water will flow, not the amount of water flowing.)
E45. Display 6.32 Two pumping systems.

a. Find the distribution of the length of time water will flow for System I.
b. Find the distribution of the length of time water will flow for System II.
c. Find the expected time that water will flow for each system and compare. Which system would you recommend?

E46. Display 6.33 gives the chances that an adult (age 18 or older) living in the United States will become a victim of the given events.

<table>
<thead>
<tr>
<th>Event</th>
<th>Rate per 1000 Adults per Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accidental injury</td>
<td>386</td>
</tr>
<tr>
<td>Personal theft</td>
<td>39</td>
</tr>
<tr>
<td>Violent victimization</td>
<td>37</td>
</tr>
<tr>
<td>Injury in motor vehicle accident</td>
<td>50</td>
</tr>
<tr>
<td>Death</td>
<td>9</td>
</tr>
<tr>
<td>Injury from fire</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Display 6.33 Rates of insurable causes of accidents or death per 1000 adults per year.

a. What should a randomly selected U.S. adult be charged yearly for a $100,000 life insurance policy if the insurance company expects to break even?
b. What should a randomly selected U.S. adult be charged yearly for a policy that pays him or her $10,000 if he or she is the victim of a violent crime?
c. An insurance company determines that people are willing to pay $50 a year for a policy that pays them a set amount if they are injured by fire. What should this set amount be if the insurance company expects to break even?

d. What is the probability that the technician will have to check at least four donations before getting the first that is type B?

E48. “One in ten high school graduates in the state of Florida sends an application to the University of Florida,” says the director of admissions there. If Florida has approximately 120,000 high school graduates next year, what are the mean and standard deviation of the number of applications the university will receive from Florida high school students? Do you see any possible weaknesses in using the binomial model here?
E49. “Girls Get Higher Grades” is a headline from USA Today. Eighty percent of the girls surveyed said it was important to them to do their best in all classes, whereas only 65% of boys responded this way. Suppose you take another random sample of 20 female students and another sample of 20 male students in your region of the country. [Source: USA Today, August 12, 1998.]

a. What is the probability that more than half the girls in the sample want to do their best in every class? What assumptions are you making?

b. What is the probability that more than half the boys in the sample want to do their best in every class? What assumptions are you making?

c. Suppose you are to take one sample of 40 students at random, instead of 20 girls and 20 boys. What is the approximate probability that more than half of those sampled will want to do their best in every class? What assumptions are you making?

E50. It is estimated that 16% of Americans have no health insurance. A polling organization randomly samples 500 Americans to ask questions about their health. [Source: U.S. Census Bureau, Current Population Survey, March 2004.]

a. What is the probability that more than 420 of those sampled will have health insurance?

b. If it costs $40 to interview a respondent with health insurance and $20 to interview a respondent without health insurance, what is the expected cost of conducting the 500 interviews?

c. If the interviews are conducted sequentially, what is the expected number of interviews to be conducted before the second person without health insurance is found?

E51. Airlines routinely overbook popular flights because they know that not all ticket holders show up. If more passengers show up than there are seats available, an airline offers passengers $100 and a seat on the next flight. A particular 120-seat commuter flight has a 10% no-show rate.

a. If the airline sells 130 tickets, how much money does it expect to have to pay out per flight?

b. How many tickets should the airline sell per flight if it wants the chance of giving any $100 payments to be about 0.05?

E52. Suppose random variable \( X \) has a geometric distribution with probability of success denoted by \( p \).

a. Find the conditional probability that \( X = 6 \) given that \( X > 5 \).

b. Find the probability that \( X > 6 \) given that \( X > 5 \).

c. Find the probability that \( X \) exceeds \( k + m \) given that it exceeds \( m \). That is, find an expression for

\[
P(X > k + m \mid X > m)
\]

This is referred to as the memoryless property of the geometric distribution.

E53. This question demonstrates one reason why the mean and variance are considered so important. Suppose you select one book at random to read from List A and one from
List B. The numbers given are the number of pages in each book.

<table>
<thead>
<tr>
<th>List A</th>
<th>List B</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>100</td>
</tr>
<tr>
<td>225</td>
<td>250</td>
</tr>
<tr>
<td>600</td>
<td>700</td>
</tr>
</tbody>
</table>

a. Find the total number of pages in each of the nine different possible pairs of books. Compute the standard deviation, $\sigma$, and variance, $\sigma^2$, of these nine totals.
b. Compute the variance of the number of pages in the three books in List A. Compute the variance of the number of pages in the three books in List B. What relationship do you see between these two variances and the variance in part a?
c. Does the same relationship as in part b hold for the three standard deviations?
d. Can you add the mean of the numbers in List A to the mean of the numbers in List B to get the mean of the nine totals?
e. Can you add the median of the numbers in List A to the median of the numbers in List B to get the median of the nine totals?

E54. Each box of a certain brand of cereal contains a poster of a famous athlete. There are five different posters, representing five different athletes.

a. Suppose you are interested in a particular athlete. Set up a simulation that would produce an approximate probability distribution of the number of boxes of cereal you would have to buy in order to get one poster of that athlete. (You stop buying cereal when you get the poster you want.)
b. What is the probability that you would get the poster of that athlete in four or fewer boxes of cereal? (You stop buying cereal when you get the poster you want.)
c. What is the expected number of boxes of cereal you would have to buy to get the one poster you want?
d. Suppose you want to get two particular posters out of the five available. What is the expected number of boxes of cereal you would have to buy to get the two specific ones? (Hint: At the outset, the probability of getting a poster you want in any one box is $\frac{2}{5}$. After you get one of them, what is the probability of getting the other poster you want? Think of the expected number of boxes of cereal being purchased in terms of these two stages.)
e. What is the expected number of boxes of cereal you would have to buy to get a full set of all five posters? Set up a simulation for this event, and compare your theoretical result to the one obtained from the simulation.

E55. Refer to the copper flipper description in D22 on page 397. This time suppose you start with 10 pennies. As in Activity 3.5a, you flip all 10, remove the ones that have “died” (landed tails), and continue flipping with a decreasing number of pennies.

a. On average, how many times do you have to flip a specified copper flipper until it “dies”?
b. If you flip 10 copper flippers at once, what is the probability that at least one “dies”?
c. On average, how many flips of the entire set of 10 copper flippers does it take to get at least one copper flipper that dies? Is the answer more or less than the answer for part a? Why?
d. Design and carry out a simulation to answer this question: On average, how many sets of flips will it take for all 10 copper flippers to die?
AP1. This table gives the percentage of women who ultimately have a given number of children. For example, 19% of women ultimately have 3 children. What is the probability that two randomly selected women will have a total of exactly 2 children?

<table>
<thead>
<tr>
<th>Number of children</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5 or more</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of women</td>
<td>18%</td>
<td>17%</td>
<td>35%</td>
<td>19%</td>
<td>7%</td>
<td>4%</td>
</tr>
</tbody>
</table>

A. 0.0289  
B. 0.1549  
C. 0.17  
D. 0.34  
E. 0.70

AP2. In a raffle, 100 tickets are sold and 10 of the tickets, selected at random, win a $5 prize. A participant buys one ticket. What is the standard deviation of the amount won?

A. $0.50  
B. $1.50  
C. $2.25  
D. $2.50  
E. $4.50

AP3. A machine producing a large spool of cable puts out 2.5 ft at a time. After each 2.5 ft of cable, there's a 1/1000 chance that the machine will cut the cable instead of continuing production. What is the expected length of the next cable the machine produces?

A. 0.0025 ft  
B. 400 ft  
C. 2,500 ft  
D. 6,250 ft  
E. 2,500,000 ft

AP4. Russell’s parents buy two apples each weekend, and put them in his lunch on two randomly selected weekdays of the following week. What is the probability that Russell gets an apple on exactly one day that is a Monday or Tuesday?

\[
\binom{5}{2} (0.4)^2 (0.6)^3 \\
2 \cdot \binom{5}{1} (0.4)^4 (0.6)^1 \\
2 \cdot \binom{5}{2} (0.4)^4 (0.6)^1 \\
\binom{6}{5} \\
\binom{2}{3} (0.2) - \binom{2}{3} (0.4)
\]

AP5. Russell’s parents roll a fair six-sided die every morning, and if the result is a 1 or 2, they pack yogurt in his lunch. What is the probability that he gets yogurt on exactly 2 of the 5 school days next week?

A. 0.111  
B. 0.400  
C. 0.165  
D. 0.329  
E. 0.671

AP6. Dawna walks 10 miles starting at 1:00 p.m. every day. On average she finishes at 3:50 p.m., with a standard deviation of 10 minutes. Jeanne leaves 30 minutes later, and every day she runs 10 miles exactly twice as fast as Dawna walks on that day. What is the mean time Jeanne will finish and the standard deviation of the number of minutes it takes Jeanne to finish?

A. mean 2:25 p.m., standard deviation 5 minutes  
B. mean 2:25 p.m., standard deviation 20 minutes
AP7. According to the World Bank, in the year 2000 approximately 35% of all people in India spent less than $1 per day. Suppose that you repeatedly take random samples of 100 Indians and record the number who spent less than $1 per day. What are the approximate mean and standard deviation of this distribution?

- mean 0.35, standard deviation 0.2275
- mean 35, standard deviation 0.2275
- mean 35, standard deviation 0.4770
- mean 35, standard deviation 4.7697
- mean 35, standard deviation 22.75

AP8. In a gas station promotion, customers receive a game piece every time they fill their tank. With each game piece, the customer has a 1 in 1250 chance of winning a $10 gasoline card. Josie decides to fill her tank at that gas station until she wins the gasoline card. What is the probability that she fills her tank there exactly 5 times?

- 0.0008
- 0.0040
- 0.9960
- 4.9800
- 7.9744

Investigative Tasks

AP9. Random sampling from a large lot of a manufactured product yields a number of defectives, X, with an approximate binomial distribution, with p being the true proportion of defectives in the lot. A sampling plan consists of specifying the number, n, of items to be sampled and an acceptance number, a. After n items are inspected, the lot is accepted if \( X \leq a \) and is rejected if \( X > a \).

a. For \( n = 5 \) and \( a = 0 \), calculate the probability of accepting the lot for values of \( p \) equal to 0, 0.1, 0.3, 0.5, and 1.

b. Graph the probability of lot acceptance as a function of \( p \) for this plan. This is the operating characteristic curve for the sampling plan.

Now, a quality-control engineer is considering two different lot acceptance sampling plans: \(( n = 5, a = 1 \) and \( n = 25, a = 5 \).

c. Construct operating characteristic curves for both plans.

d. If you were a seller producing lots with proportions of defectives between 0% and 10%, which plan would you prefer?

e. If you were a buyer wishing high protection against accepting lots with proportions of defectives over 30%, which plan would you prefer?
What would happen if you could take random samples over and over again from your population? Sampling distributions show how much your results might vary from sample to sample, as when estimating the mean number of departures from U.S. airports for a given period of time.
You have studied methods that are good for exploration and description, but for inference—going beyond the data in hand to conclusions about the population or the process that created the data—you need to collect the data by using a random sample (a survey) or by randomly assigning treatments to subjects (an experiment). The promise of Chapter 4 was that if you used those methods to produce a data set, you then could use your data to draw sound conclusions. Randomized data production not only protects against bias and confounding but also makes it possible to imagine repeating the data production process so you can estimate how the summary statistic you compute from the data would vary from sample to sample. To oversimplify, but only a little, a sampling distribution is what you get by repeating the process of producing the data and computing the summary statistic many times.

**In this chapter, you will learn**

- how to use simulation to generate approximate sampling distributions of common summary statistics such as the sample mean and the sample proportion
- to describe the shape, center, and spread of the sampling distributions of common summary statistics without actually generating them
- to use the sampling distribution to determine which results are reasonably likely and which would be considered rare
Generating Sampling Distributions

Sometimes it is possible to imagine repeating the process of data collection. A **sampling distribution** is the distribution of the summary statistics you get from taking repeated random samples. For example, in Chapter 1 (Display 1.10), you created a simulated sampling distribution of the mean age from random samples of three employees who could have been laid off at Westvaco. This distribution is shown again in Display 7.1. There were ten workers who could have been laid off; their ages were

```
25 33 35 38 48 55 55 55 56 64
```

You randomly selected three of these workers—without replacement, because Westvaco couldn’t have laid off the same worker twice.

Your summary statistic was the mean (average) age of the three workers you chose at random. You were able to repeat this process many times in order to generate a distribution of possible values of the summary statistic.

Using this simulated sampling distribution, you could make a decision about whether it was reasonable to assume that employees were selected for layoff without respect to their age. The ages of the three workers actually laid off were 55, 55, and 64, for an average age of 58. As you can see from Display 7.1, it’s rather hard to get an average age that large just by chance. So Westvaco had some explaining to do.

![Display 7.1](image)

**Display 7.1** A simulated sampling distribution of the mean age from random samples of three people who could have been laid off at Westvaco.

In the Westvaco analysis, you went through four steps in using simulation to generate an approximate sampling distribution of the mean age of three workers:

1. Take a random sample of a fixed size $n$ from a population.
2. Compute a summary statistic.
3. Repeat steps 1 and 2 many times.
4. Display the distribution of the summary statistics.
These steps describe how to use simulation to make an *approximate sampling distribution or simulated sampling distribution*. If you had listed all possible samples of 3 workers who could have been laid off from the 10 workers and computed the mean age for all \( \binom{10}{3} = 120 \) samples, you would have constructed the exact or theoretical sampling distribution, called simply the *sampling distribution*.

**Shape, Center, and Spread: Now and Forever**

In Chapter 4, you sampled from a population of 100 rectangles. Display 7.2 shows their areas, which have mean 7.4 and standard deviation 5.2.

Display 7.2  The population of rectangle areas.

In Activity 4.2a on pages 232–235, you generated an approximate sampling distribution when you selected five rectangles at random and calculated their mean area, doing this over and over again. You probably got a distribution something like that in Display 7.3. Each of the 1000 dots in Display 7.3 represents the mean area of a random sample of five rectangles.

Display 7.3  A simulated sampling distribution of the sample mean of five rectangles.

A good description of a sampling distribution like that in Display 7.3 is still the trio

Shape  Center  Spread
This simulated sampling distribution is approximately normal, with a hint of a skew to the right, with mean 7.4 and standard deviation 2.3.

Compare Display 7.3 with the population of the areas of all 100 rectangles in Display 7.2. The two distributions both have mean 7.4. However, there are two quite different standard deviations in the picture—the standard deviation of the population, 5.2, and the smaller standard deviation of the sampling distribution, 2.3. To tell them apart, these two types of standard deviations have different names. The standard deviation of the population, \( \sigma \), is called, naturally enough, the population standard deviation. The standard deviation of the sampling distribution is commonly called the standard error, or SE for short.

If you select one of the rectangles at random and get one with an area of 16, you shouldn’t be at all surprised. You can see from Display 7.2 that 15 of the 100 rectangles, or 15%, have an area of 16 or larger. Values like this that lie in the middle 95% of a sampling distribution are called reasonably likely events. Values that lie in the outer 5% of a sampling distribution are called rare events. In a normal distribution, rare events are those values that lie more than two standard deviations from the mean.

Notice that if you take a random sample of five rectangles and their mean area is 16, you should be very surprised. You can see from the simulation in Display 7.3 that a mean this large happened only once in 1000 runs and so would be a rare event.

**DISCUSSION**

**Shape, Center, and Spread**

D1. Compare Displays 7.2 and 7.3.

a. Why is the largest value in Display 7.2 larger than the largest value in Display 7.3? Why is the smallest value smaller?

b. Would you expect the means of the distributions in Displays 7.2 and 7.3 to be equal? Why or why not?

c. Would you expect the two standard deviations to be equal? Why or why not?

d. How does the shape of the sampling distribution for samples of size 5 compare to the shape of the distribution of the population of the areas of all 100 rectangles?

D2. Approximately what values of the sample mean for samples of size 5 would be reasonably likely? Which would be rare events?

Once you’re used to the steps involved in generating sampling distributions and have a calculator or computer that lets you do simulations easily, there is no end to the variety of summary statistics you can study. [See Calculator Note 7A to learn about the kinds of sampling distributions you can generate with a calculator.]

In Activity 7.1a, you will use the 100 rectangles from Display 4.5 on page 233 to generate simulated sampling distributions of two summary statistics other than the mean: the sample median and the sample maximum.
ACTIVITY 7.1a

The Return of the Random Rectangles

**What you'll need:** a copy of Display 4.5, a method of producing random digits

1. Generate five distinct random numbers between 00 and 99.
2. Find the rectangles in Display 4.5 on page 233 that correspond to your random numbers. (The rectangle numbered 100 can be called 00.) This is your random sample of five rectangles.
3. Determine the areas of the rectangles in your random sample and find the sample median. (Save these five areas for step 7.)
4. Combine your sample median with those of other students in the class and repeat steps 1 through 3 until you have 200 sample medians. Construct a plot showing the distribution of the 200 sample medians.
5. Describe the shape, mean, and standard error of this plot of sample medians. How do these compare to the shape, mean, and standard deviation of the population of areas in Display 7.2?
6. How does this sampling distribution compare to the simulated sampling distribution of the sample mean in Display 7.3?
7. Repeat steps 4 and 5, but this time use the maximum area of the five rectangles in each sample as the summary statistic. For example, if the areas in your sample are 1, 8, 4, 8, and 3, the maximum area is 8. (You can use the same samples as before.) You will also need to keep your data for E13.
8. Compare the population in Display 7.2 to the sampling distributions you plotted.
   a. What is the median of the population of 100 rectangle areas? Is the sample median a good estimate of the median of the areas of all 100 rectangles?
   b. What is the maximum of the population of 100 rectangle areas? Is the sample maximum a good estimate of the largest area of the 100 rectangles?
9. **Reasonably likely and rare events.** Use the plots you constructed to answer these questions.
   a. What values of the sample median are reasonably likely? What values would be rare events?
   b. What values of the sample maximum are reasonably likely? What values would be rare events?

**Exact Sampling Distributions**

When the population size is very small, you can construct sampling distributions exactly by listing all possible samples.

**Example: Mapping National Parks**

Utah has five national parks. Your company has been hired to make maps of two of these parks, which will be selected at random. Use the information in Display 7.4 to construct the sampling distribution for the total number of square
miles you would map. Then find the probability that you will have to map more than 600 square miles.

<table>
<thead>
<tr>
<th>National Park</th>
<th>Area (sq mi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arches (A)</td>
<td>119</td>
</tr>
<tr>
<td>Bryce Canyon (B)</td>
<td>56</td>
</tr>
<tr>
<td>Canyonlands (C)</td>
<td>527</td>
</tr>
<tr>
<td>Capitol Reef (R)</td>
<td>378</td>
</tr>
<tr>
<td>Zion (Z)</td>
<td>229</td>
</tr>
</tbody>
</table>

Display 7.4 Sizes of Utah’s national parks.

**Solution**

There are \( \binom{5}{2} \), or 10, equally likely ways to select two national parks, as shown in Display 7.5. The relative frequency histogram shows the sampling distribution of the sum. The probability that you will have to map more than 600 square miles is \( \frac{4}{10} \).
Properties of Point Estimators

Sampling distributions are the connecting link between the collection of data through designed studies based on random sampling or random assignment of treatments to subjects, and statistical inference, the process of drawing defensible conclusions from the data. When inferences are made from the sample to the population, the sample mean is viewed as an estimator of the mean of the population from which the sample was selected. Similarly, the proportion of successes in a sample is an estimator of the proportion of successes in the population. These two statistics are called point estimators for the reason that, for a given sample, they provide only one point (number) as a plausible value of the corresponding population parameter. (In the coming chapters you will learn how to construct an interval of plausible values of the population parameter.)

Suppose you want to know the mean area of the rectangles whose areas are plotted in Display 7.2. Because of time or money constraints, you cannot measure them all. You know what to do: Take a random sample from the population of rectangles; compute the sample mean area, \( \bar{x} \); use \( \bar{x} \) as your estimator of the population mean area, \( \mu \).

This process seems so natural that it is easy to overlook just how powerful it is. Two very fortunate things are happening in the background:

- Although each individual sample mean \( \bar{x} \) usually is smaller or larger than the population mean \( \mu \), in the long run, over many samples, the average value of \( \bar{x} \) turns out to be exactly equal to \( \mu \).
- The value of \( \bar{x} \) tends to be pretty close to the value of \( \mu \), and, as you will see in the next section, the larger the sample size, the closer to \( \mu \) it tends to be.

Even though in practice you can’t repeat the sampling process, it is nice to know whether the estimator you are using generally works out well. For example, from Display 7.3, you can see that while a single estimate \( \bar{x} \) from a sample of size 5 might be too small or too large, the estimates are centered around the population mean, 7.4. Display 7.3 also gives you an idea of how close your estimate is likely to be to 7.4 if you sample only five rectangles. Some desirable properties of estimators are generalized in the box.

### Properties of Point Estimators

When you use a summary statistic from a sample to estimate a parameter of a population, there are two properties that you would like the summary statistic to have.

- The summary statistic should be **unbiased**. That is, the mean of the sampling distribution is equal to the value you would get if you computed the summary statistic using the entire population. More formally, an estimator is unbiased if its expected value equals the parameter being estimated.
- The summary statistic should have as little variability as possible (be more **precise** than other estimates) and should have a standard error that decreases as the sample size increases.
**Example: Estimating the Distance of the Farthest Galaxy**

Your observatory is given the names of 31 galaxies. Your job is to estimate the maximum distance from Earth among the 31 galaxies. You have the resources to measure the distance of only 10 galaxies, which you will select at random. Unbeknownst to you, the distances, in megaparsecs (Mpc), are

0.008 0.76 0.81 7.2 7.5 11.4 11.7 13.2 15 15 15.3 15.5 15.7 16.1 16.8 16.8 20.9 22.9 22.9 24.1 25.9 26.2 29.2 31.6 58.7 93

The maximum distance, 93 Mpc, is the population parameter you hope to estimate.

Because you want to estimate the maximum distance from Earth, you decide to use the maximum distance in the sample as your summary statistic. For example, if your sample of size 10 is

0.81 7.5 11.6 11.7 13.2 16.1 16.8 24.1 26.2 58.7

you will estimate the population maximum to be 58.7. This is smaller than the maximum in the population, so you have underestimated.

Display 7.6 shows a simulated sampling distribution of the maximum of samples of size 10. What will happen when you use the maximum of the sample as an estimator of the maximum of this population?

**Display 7.6** Simulated sampling distribution of the maximum of samples of size 10 taken from the population of 31 galaxies.

**Solution**

The population maximum is 93, but the mean of the sampling distribution is only 56. If you could repeat your process of estimating the population maximum by using the maximum in the sample, on average your estimate would be too small. In other words, the sample maximum is a biased estimator of the population maximum. That’s not too surprising, because the maximum of a sample can never be larger than the maximum of its population.
Further, there is a great deal of variability in the sample maximum. Your estimate of the population maximum might be as high as 93 Mpc (the population maximum), but about half the time your estimate will be less than 32, way too small. Using the sample maximum as an estimator of the population maximum doesn’t work out very well.

**Properties of Point Estimators**

D3. In what way are the ideas “biased estimator” and “bias in sampling” similar? In what way are they different?

**Summary 7.1: Sampling Distributions**

A sampling distribution answers the question “How would my summary statistic behave if I could repeat the process of collecting data using random samples?” By showing you the summary statistics from random samples, a sampling distribution can help you decide whether the same process created the result you got from your actual sample.

You can generate a simulated sampling distribution of any sample statistic by following these steps:

1. Take a random sample of size $n$ from a population.
2. Compute a summary statistic.
3. Repeat steps 1 and 2 many times.
4. Display the distribution of the summary statistics.

You can sometimes create an exact sampling distribution by listing all possible samples. In many more situations, you will be able to describe the shape, mean, and standard error of the sampling distribution without simulation and without listing samples. You will learn how to do this for sample means and sample proportions in Sections 7.2 and 7.3.

Here are several key facts about sampling distributions:

- A simulated, or approximate, sampling distribution is the distribution of the sample statistic for a large number of repeated random samples.
- Sampling distributions, like data distributions, are best described by shape, center, and spread.
- The standard deviation of a sampling distribution is called the standard error.
- Many, but not all, sampling distributions are approximately normal. For normal distributions, reasonably likely outcomes are those that fall within approximately two standard errors of the mean.
- If the mean of the sampling distribution is equal to the population parameter being estimated, then the summary statistic you are using is an unbiased estimator of that parameter.
Practice
Shape, Center, and Spread

P1. Display 7.7 gives the number of moons for each planet in our solar system. For a term paper in your astronomy class, you will be given three different planets selected at random and be asked to write a 50-word description of each of the planets’ moons.

<table>
<thead>
<tr>
<th>Planet</th>
<th>Number of Moons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mercury</td>
<td>0</td>
</tr>
<tr>
<td>Venus</td>
<td>0</td>
</tr>
<tr>
<td>Earth</td>
<td>1</td>
</tr>
<tr>
<td>Mars</td>
<td>2</td>
</tr>
<tr>
<td>Jupiter</td>
<td>63</td>
</tr>
<tr>
<td>Saturn</td>
<td>56</td>
</tr>
<tr>
<td>Uranus</td>
<td>27</td>
</tr>
<tr>
<td>Neptune</td>
<td>13</td>
</tr>
</tbody>
</table>

Display 7.7 Number of moons for the planets in our solar system. [Source: NASA, solarsystem.nasa.gov.]

a. What is the smallest number of words you might have to write? The largest?
b. Describe how to generate a simulated sampling distribution of the total number of moons you must describe.
c. Generate 20 values for a simulated sampling distribution and make a plot.
d. Use your plot from part c to estimate the probability that you will have to write 150 words or fewer.

Jupiter and its four planet-sized moons

P2. Display 7.8 shows the distribution of exam scores for 192 students in an introductory statistics course at the University of Florida.

Display 7.8 A distribution of exam scores.

a. Match each histogram in Display 7.9 to its description.
   I. the individual scores for one random sample of 30 students
   II. a simulated sampling distribution of the mean of the scores of 100 random samples of 4 students
   III. a simulated sampling distribution of the mean of the scores of 100 random samples of 30 students

b. The second-hour class of 30 students, whose exam scores are part of the data set, had a class average of 86 on this exam. The instructor says that this class is just a random assortment of typical students taking this course. Do you agree?
P3. Using the exam scores of P2, simulated sampling distributions were generated for the minimum, maximum, median, and lower quartile for samples of size 10.
   a. Match each histogram in Display 7.10 to its summary statistic.
      I. the minimum
      II. the maximum
      III. the median
      IV. the lower quartile
   b. Which sampling distribution has the largest standard error? The next largest?

Display 7.10  Histograms of different summary statistics from random samples of size 10 of statistics exam scores.

Exact Sampling Distributions

P4. Every year, Forbes magazine releases a list of the top-earning dead celebrities. In 2005, the top six and their yearly earnings were
   1. Elvis Presley, $45 million
   2. Charles M. Schulz, $35 million
   3. John Lennon, $22 million
   4. Andy Warhol, $16 million
   5. Theodor “Dr. Seuss” Geisel, $10 million
   6. Marlon Brando, $9 million
   [Source: www.forbes.com.]

Your talent agency gets an opportunity to represent two of these dead celebrities, to be selected at random. You will be paid 10% of their earnings.
   a. What is the most you could be paid? The least?
   b. Construct the sampling distribution of your total possible earnings.
   c. What is the probability that you will be paid $3 million or more?

Properties of Point Estimators

P5. The areas of the five national parks in Utah are given in Display 7.11.

<table>
<thead>
<tr>
<th>National Park</th>
<th>Area (sq mi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arches (A)</td>
<td>119</td>
</tr>
<tr>
<td>Bryce Canyon (B)</td>
<td>56</td>
</tr>
<tr>
<td>Canyonlands (C)</td>
<td>527</td>
</tr>
<tr>
<td>Capitol Reef (R)</td>
<td>378</td>
</tr>
<tr>
<td>Zion (Z)</td>
<td>229</td>
</tr>
</tbody>
</table>

Display 7.11  Areas of the five national parks in Utah.

   a. What is the mean area of this population of parks? The standard deviation?
   b. Complete the table in Display 7.12 (on the next page) and use it to construct the sampling distribution of the mean area of a random sample of two parks.
Sample of Two Parks  | Total Area (sq mi) | Mean Area (sq mi) |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A and B</td>
<td>175</td>
<td>87.5</td>
</tr>
<tr>
<td>A and C</td>
<td>646</td>
<td>323</td>
</tr>
<tr>
<td>A and R</td>
<td>497</td>
<td>—?—</td>
</tr>
<tr>
<td>A and Z</td>
<td>348</td>
<td>—?—</td>
</tr>
<tr>
<td>B and C</td>
<td>583</td>
<td>—?—</td>
</tr>
<tr>
<td>B and R</td>
<td>434</td>
<td>—?—</td>
</tr>
<tr>
<td>B and Z</td>
<td>285</td>
<td>—?—</td>
</tr>
<tr>
<td>C and R</td>
<td>905</td>
<td>—?—</td>
</tr>
<tr>
<td>C and Z</td>
<td>756</td>
<td>—?—</td>
</tr>
<tr>
<td>R and Z</td>
<td>607</td>
<td>—?—</td>
</tr>
</tbody>
</table>

Display 7.12 Mean area of two randomly chosen national parks in Utah.

c. Verify that the sample mean is an unbiased estimator of the population mean. (In other words, verify that the mean of the sampling distribution is exactly equal to the mean of the population.)

d. Verify that the standard error of the sampling distribution is smaller than the standard deviation of the population.

P6. Refer to the example on page 413. Your company’s contract has changed. You now are to estimate the range (largest minus smallest) of the areas of Utah’s national parks. Your procedure will be to select three of the five parks at random and use the range of their three areas as the estimate.

<table>
<thead>
<tr>
<th>Sample of Three Parks</th>
<th>Range of the Areas (sq mi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A, B, C</td>
<td>527 − 56 = 471</td>
</tr>
<tr>
<td>A, B, R</td>
<td>378 − 56 = 322</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Display 7.13 Range of areas of three randomly chosen national parks in Utah.

Exercises

E1. Random samples are taken from the population of random digits 0 through 9, with replacement.

a. Each histogram in Display 7.14 is a simulated sampling distribution of the sample mean. Match each sampling distribution to the sample size used: 1, 2, 20, or 50.

b. Compare the means. How does the mean of the sampling distribution depend on the sample size?

c. Compare the spreads. How does the spread of the sampling distribution depend on the sample size?

E2. Three very small populations are given, each with a mean of 30.

A. 10 50
B. 10 20 30 40 50
C. 20 30 40

Match each population to the sampling distribution of the sample mean (Display 7.15) for a sample of size 2 (taken with replacement).
E3. The mean and the median are only two of many possible measures of center. Another is the **midrange**, defined as the midpoint between the minimum and the maximum in a data set. The rectangles in Display 4.5 on page 233 have areas ranging from 1 to 18, so the population midrange is \( \frac{1 + 18}{2} \), or 9.5. The histogram in Display 7.16 shows a simulated sampling distribution of the midrange based on 1000 random samples of size five from the population of 100 rectangles. The mean of this distribution is 8.05, and the estimated standard error is 2.17.

a. Take a random sample of five rectangles; compute the midrange of their areas.

b. Describe the sampling distribution of the midrange. How does it compare to the sampling distribution of the sample mean in Display 7.3 on page 411?

c. Would you use the midrange as a measure of center for the rectangle area population? Why or why not?
E4. The areas of all 50 U.S. states, in millions of acres, are given in Display 7.17.

<table>
<thead>
<tr>
<th>State</th>
<th>Acres (in millions)</th>
<th>State</th>
<th>Acres (in millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL</td>
<td>33.1</td>
<td>MT</td>
<td>94.1</td>
</tr>
<tr>
<td>AK</td>
<td>366.0</td>
<td>NC</td>
<td>33.7</td>
</tr>
<tr>
<td>AR</td>
<td>34.0</td>
<td>ND</td>
<td>45.2</td>
</tr>
<tr>
<td>AZ</td>
<td>73.0</td>
<td>NE</td>
<td>49.5</td>
</tr>
<tr>
<td>CA</td>
<td>101.6</td>
<td>NH</td>
<td>5.9</td>
</tr>
<tr>
<td>CO</td>
<td>66.6</td>
<td>NJ</td>
<td>5.0</td>
</tr>
<tr>
<td>CT</td>
<td>3.2</td>
<td>NM</td>
<td>77.8</td>
</tr>
<tr>
<td>DE</td>
<td>1.3</td>
<td>NV</td>
<td>70.8</td>
</tr>
<tr>
<td>FL</td>
<td>37.5</td>
<td>NY</td>
<td>31.4</td>
</tr>
<tr>
<td>GA</td>
<td>37.7</td>
<td>OH</td>
<td>26.4</td>
</tr>
<tr>
<td>HI</td>
<td>4.1</td>
<td>OK</td>
<td>44.8</td>
</tr>
<tr>
<td>IA</td>
<td>36.0</td>
<td>OR</td>
<td>62.1</td>
</tr>
<tr>
<td>ID</td>
<td>53.5</td>
<td>PA</td>
<td>29.0</td>
</tr>
<tr>
<td>IL</td>
<td>36.1</td>
<td>RI</td>
<td>0.8</td>
</tr>
<tr>
<td>IN</td>
<td>23.1</td>
<td>SC</td>
<td>19.9</td>
</tr>
<tr>
<td>KS</td>
<td>52.6</td>
<td>SD</td>
<td>49.4</td>
</tr>
<tr>
<td>KY</td>
<td>25.9</td>
<td>TN</td>
<td>27.0</td>
</tr>
<tr>
<td>LA</td>
<td>30.6</td>
<td>TX</td>
<td>170.8</td>
</tr>
<tr>
<td>MA</td>
<td>5.3</td>
<td>UT</td>
<td>54.3</td>
</tr>
<tr>
<td>MD</td>
<td>6.7</td>
<td>VA</td>
<td>26.1</td>
</tr>
<tr>
<td>ME</td>
<td>21.3</td>
<td>VT</td>
<td>6.1</td>
</tr>
<tr>
<td>MI</td>
<td>37.4</td>
<td>WA</td>
<td>43.6</td>
</tr>
<tr>
<td>MN</td>
<td>54.0</td>
<td>WI</td>
<td>35.9</td>
</tr>
<tr>
<td>MO</td>
<td>44.6</td>
<td>WV</td>
<td>15.5</td>
</tr>
<tr>
<td>MS</td>
<td>30.5</td>
<td>WY</td>
<td>62.6</td>
</tr>
</tbody>
</table>

a. Describe the shape of the distribution.

b. Use a table of random digits to select a random sample of five areas. Do these areas appear to be representative of the population? What is the mean of your five areas?

c. The plot in Display 7.18 shows the mean area for 25 random samples, each of size 5, just like the sample you generated in part b. Describe this simulated sampling distribution.

d. From what particular sample could the largest value in the plot in Display 7.18 have come?

E5. The five tennis balls in a can have diameters 62, 63, 64, 64, and 65 mm. Suppose you select two of the tennis balls at random, without replacing the first before selecting the second.

a. Construct a dot plot of the five population values.

b. List all possible sets of size 2 that can be chosen from the five balls. There are \( _5 \text{C}_2 \) (“5 choose 2”), or 10, possible sets of two balls.

c. List all 10 possible sample means, and construct a dot plot of the sampling distribution of the sample mean. Compute the mean and standard error of this distribution. Compare these to the mean and standard deviation of the population.
d. Construct a dot plot of the sampling distribution of the maximum diameter of a sample of size 2.

e. Construct a dot plot of the sampling distribution of the range (maximum minus minimum) of a sample of size 2.

E6. As part of a statistics project at Iowa State University, a student tested how well a bike with treaded tires stopped on concrete. In six trials, these lengths of skid marks (in centimeters) were produced: 365, 374, 376, 391, 401, 402. Suppose instead that the student had done only three trials, which could have been any three selected from this population of six, each with equal probability of being selected. [Source: Stephen B. Vardeman, *Statistics for Engineering Problem Solving* (Boston: Prindle, Weber & Schmidt, 1994), p. 349.]

a. Construct a dot plot of the six population values.

b. List all possible sets of three skid lengths the student could get. There are $\binom{6}{3}$, or 20, possible sets.

c. Construct a dot plot of the sampling distribution of the sample mean. Find its mean and standard error. Is the mean equal to the mean of the population?

d. Construct a dot plot of the sampling distribution of the sample median. Find its median. Is it equal to the median of the population?

e. Construct a dot plot of the sampling distribution of the sample minimum. Describe the shape of this distribution. Is the sample minimum a good estimator of the population minimum of all six lengths? Explain.

E7. Refer to your exact sampling distributions in E5. For each summary statistic here, explain why the summary statistic is or is not an unbiased estimator of the population parameter. If it is a biased estimator, does it tend to be too big or too small, on average?

a. sample mean (see part c)

b. sample maximum (see part d)

c. sample range (see part e)

E8. Refer to your exact sampling distributions in E6. For each summary statistic here, explain why the summary statistic is or is not an unbiased estimator of the population parameter. If it is a biased estimator, does it tend to be too big or too small, on average?

a. sample mean (see part c)

b. sample median (see part d)

c. sample minimum (see part e)

E9. An inspector is called to see how much damage has been caused to a warehouse full of frozen fish by a recent power failure. The fish are stored in cartons containing 24 one-pound packages. There are hundreds of cartons in the warehouse.

The inspector decides he needs a sample of 48 one-pound packages of fish in order to assess the damage. He selects two cartons at random and treats this as a sample of 48 one-pound units.

The inspector doesn’t realize it, but if one package of fish in a carton spoils, very rapidly the whole carton spoils. So it is safe to assume that each carton in the warehouse is either completely spoiled or not spoiled at all.

a. If the inspector is using pounds of spoiled fish in the sample as a statistic, what will the sampling distribution of the statistic look like? Describe this distribution as completely as you can.

b. Did the inspector choose a good sampling plan? If not, explain how he could have chosen a better one.
E10. Joel knows that the sample range is a biased estimator of the population range, so he has a plan to use a different estimator. Instead of using the sample range to estimate the population range, he will double the sample interquartile range (IQR) and use that as an estimate of the population range. He will try this out using the population of numbers \{0, 1, 2, 3, \ldots, 98, 99, 100\} and a sample size of 8, taken with replacement.

a. What is the value of the population parameter Joel is trying to estimate?

b. Explain why the sample range is a biased estimator of the population range. You can use Display 7.19 in your explanation.

c. To test his estimator, Joel does a simulation. He takes 10,000 samples of size 8, computes the IQR of each one, and doubles it. His results are shown in Display 7.20. Does Joel’s estimator of the range appear to be biased? Do you have any other concerns about it?

For E11–E12: In Chapter 2, you might have wondered why the denominator of the sample standard deviation is \( n - 1 \) rather than \( n \). Now that you know about sampling distributions, you can use them to answer this question in true statistical fashion. If a sample standard deviation is to be a good estimator of the population standard
deviation $\sigma$, its sampling distribution should be centered at $\sigma$, or at least near that value. In E11 and E12, you will see whether that turns out to be the case.

E11. Suppose you have a population that consists of only the three numbers 2, 4, and 6. You take a sample of size 2, replacing the first number before you select the second. Use parts a–e to complete the table in Display 7.21.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Mean</th>
<th>Variance, Dividing by $n = 2$</th>
<th>Variance, Dividing by $n - 1 = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2, 2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2, 4</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Average</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Display 7.21 Variance calculations.

a. In the first column, list all nine possible samples. (Count 2, 4 as different from 4, 2; that is, order matters.)

b. Compute the mean of each sample and write it in the second column.

c. Compute the variance of each sample, dividing by $n = 2$, and enter it in the third column. (You should be able to do this in your head.)

d. Compute the variance of each sample, dividing by $n - 1$, or 1, and enter it in the fourth column.

e. Compute the average of each column.

f. Compute the variance of the population {2, 4, 6} using the formula on page 366 with the probabilities of selection being equal for the three values. Compare it to your answers in part e.

g. Do you get an unbiased estimate of the variance of the population if you divide by $n$ or by $n - 1$? Explain how you know.

h. If you divide by $n$ rather than by $n - 1$, does the variance tend to be too large or too small, on average?

E12. Look at the rectangle areas plotted in Display 7.2 on page 411. Take a random sample of five rectangles. Compute the standard deviation of the areas of these rectangles using the regular formula, dividing by $n - 1$, or 4. Then compute the standard deviation again, but divide by $n$, or 5. Which standard deviation is larger? The mean area of the population is 7.42, and the population standard deviation, $\sigma$, is 5.20. The first histogram in Display 7.22 shows the sample standard deviations from the same samples of size 5 whose means are plotted in Display 7.3 (page 411). The mean of these standard deviations is 5.02.

If you divide by 5 rather than by 4 (that is, by $n$ instead of $n - 1$), the results produce the distribution shown in the second histogram. The mean of this distribution is 4.49. Explain why dividing by $n - 1$ gives a better estimate of the population standard deviation.

Display 7.22 Simulated sampling distributions of sample standard deviations ($n = 5$).
E13. As you found in Activity 7.1a, step 7, if you use the sample maximum as an estimate of the population maximum, it tends to be too small. Here is one possible rule for adjusting it:

\[
\text{(population maximum)} = \frac{n + 1}{n} \cdot \text{(sample maximum)}
\]

As always, \( n \) is the sample size. This rule works best when the population is uniform.

a. Suppose you take a random sample of four numbers from the uniform distribution on \([0, N]\), where \( N \) is the population maximum. On average, what size gap would you expect between two adjacent numbers in your sample? The dot plot in Display 7.23 might help. It shows a sample of size 4 taken from a uniform distribution with minimum 0 and maximum \( N \).

![Display 7.23](image)

b. Where would you expect the maximum of a sample of size 4 to be?

c. If you know the sample maximum of a sample of size 4 taken from a uniform distribution on \([0, N]\), what would you expect \( N \) to be?

d. Test how well this rule works using your results from Activity 7.1a, step 7.

E14. You have seen a number of examples in which the variation in the sampling distribution of the sample mean is smaller than the variation in the sampling distribution of the sample median. However, these examples arose from populations that were symmetric, or nearly so. Suppose the distribution for your population is highly skewed (for example, the areas of the U.S. states given in E4). Will the sampling distribution of the sample median have less variation than the sampling distribution of the sample mean? Use simulation to provide evidence for your answer.

---

### 7.2 Sampling Distribution of the Sample Mean

You have seen that the sample mean (average) is a very important statistic, especially for symmetric distributions. In fact, you knew that the mean is important before you ever took a statistics course because you see averages used all around you—average exam score, average income, average age, batting average, and on and on. In this section, you will learn to predict the mean, standard deviation, and shape of the sampling distribution of the sample mean. Because its properties are so well understood, the sample mean will be used almost exclusively as the measure of center in the remainder of this book.

Some of the distributions of data that you have studied so far have had a roughly normal shape, but many others were not at all normal. Part of what makes the normal distribution important is its tendency to emerge when you create sampling distributions. Activity 7.2a shows how the normal distribution comes up unexpectedly when you work with the sampling distribution of a mean.
**ACTIVITY 7.2a**

**Cents and Center**

**What you’ll need:** 25 pennies collected from recent day-to-day change

1. If you were to construct a histogram of the ages of all the pennies from all the students in your class, what do you think the shape of the distribution would look like?

2. Find the age of each of your pennies by subtracting the date on the penny from the current year. Construct a histogram of the ages of all the pennies in the class.

3. Estimate the mean and standard deviation of the distribution. Then confirm these estimates by actual computation.

4. Take a random sample of size 5 from the ages of your class’s pennies, and compute the mean age of your sample.

5. If you were to construct a histogram of all the mean ages computed by the students in your class in step 4, do you think the mean of the values in this histogram would be larger than, smaller than, or the same as the mean for the population of the ages of all the pennies? Regardless of your choice, try to make an argument to support each choice. Estimate what the standard deviation of the distribution of the mean ages will be.

6. Construct the histogram of the mean ages, using nickels if you have them, and determine the mean and standard deviation. Which of the three choices in step 5 appears to be correct?

7. Repeat steps 4–6 for samples of size 10, using dimes, and size 25, using quarters.

8. Look at the four histograms that your class has constructed. What can you say about the shape of the histogram as \( n \) increases? About the center? About the spread?

In the rest of this section, you’ll see if the results from Activity 7.2a, step 8, are true for other distributions.

**Shape, Center, and Spread of the Sampling Distribution of \( \bar{x} \)**

Display 7.24 shows the distribution of the number of children per family in the United States. This isn’t a sampling distribution; rather, it’s the population distribution you will use to build a sampling distribution.

If you count all families with four or more children as having four children, this highly skewed population has a mean of about 0.9 and a standard deviation of about 1.1. [See Calculator Note 6A to review how to calculate the mean and standard deviation of a probability distribution.]

Suppose you are working for a video game company that wants to sample families to study interests of children. What will the sampling distributions of the mean number of children per family look like? These four steps review how to construct a simulated sampling distribution of the mean for samples of size 4.

1. Take a random sample from a population.

In your model, there should be a 0.524 chance that a randomly selected family will have no children, and so on. Use a table of random digits, as you did in Section 5.2, or use a calculator’s random number generator to select a random sample of four families from this distribution.
2. Compute a summary statistic for your random sample.

Suppose your random sample represents families with two, two, one, and two children. The sample mean is then

\[ \bar{x} = \frac{(2 + 2 + 1 + 2)}{4} = 1.75 \]

3. Repeat steps 1 and 2 many times.

4. Display the distribution of the summary statistics.

The five sampling distributions of the sample mean in Display 7.25 are for samples of size 1, 4, 10, 20, and 40. Note that the exact sampling distribution of the mean for samples of size 1 is identical to the population distribution.

Display 7.25  Sampling distributions of the sample mean for samples of size 1, 4, 10, 20, and 40.
The plots and summary table of these sampling distributions reveal some interesting patterns:

- The population started out with only five different values and a large skew toward the larger numbers. The sampling distribution for samples of size 1 looks almost identical to the distribution of the population, as it should. The sampling distribution for samples of size 4 is still somewhat skewed but has more values. The sampling distribution for samples of size 10 is even less skewed. The sampling distributions for samples of size 20 and 40 look nearly normal. (This progression toward normality will turn out to be very important.)

- The means of all five sampling distributions are equal, 0.9. This is the mean of the population.

- The standard errors of the five sampling distributions decrease from 1.1 for samples of size 1 (which is the same as the population standard deviation) to 0.55 for samples of size 4, to 0.35 for samples of size 10, and so on. Just as you would predict, the sample mean tends to be closer to the population mean with larger samples than it is with smaller samples.

It is time to distill the information you have worked with into a set of rules for describing the behavior of sampling distributions of the sample mean. The symbols commonly used are organized in this table.

<table>
<thead>
<tr>
<th>Common symbols</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Population Parameter</th>
<th>Sample Statistic</th>
<th>Sampling Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>( \mu )</td>
<td>( \mu )</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>( \sigma )</td>
<td>( s )</td>
</tr>
<tr>
<td>Size</td>
<td>( N )</td>
<td>( n )</td>
</tr>
</tbody>
</table>

### Properties of the Sampling Distribution of the Sample Mean

If a random sample of size \( n \) is selected from a population with mean \( \mu \) and standard deviation \( \sigma \), then

**Center**

- the mean, \( \mu_x \), of the sampling distribution of \( \bar{x} \) equals the mean of the population, \( \mu \):

  \[
  \mu_x = \mu
  \]

**Spread**

- the standard deviation, \( \sigma_x \), of the sampling distribution of \( \bar{x} \), sometimes called the **standard error of the mean**, equals the standard deviation of the population, \( \sigma \), divided by the square root of the sample size \( n \):

  \[
  \sigma_x = \frac{\sigma}{\sqrt{n}}
  \]

**Shape**

- the shape of the sampling distribution will be approximately normal if the population is approximately normal; for other populations, the sampling distribution becomes more normal as \( n \) increases (This property is called the **Central Limit Theorem**.)
All three properties are of great importance in statistics, and all three depend on random samples. The first property, \( \mu_x = \mu \), says that the means of random samples are centered at the population mean. This might seem obvious, and it is obvious for symmetric populations. But as you saw in Activity 7.2a, it’s also true for skewed populations.

The second property, \( \sigma_x = \sigma / \sqrt{n} \), is the main reason for using the standard deviation to measure spread—you can find the standard error of the sample mean without simulation. This result validates our intuitive feeling of why large samples are better: The larger the sample, the closer its mean tends to be to the population mean.

The third property, the Central Limit Theorem, deals with the shape of the sampling distribution and will be central to your work in the rest of this course. It helps you decide which outcomes are reasonably likely and which are not.

**DISCUSSION**

**Shape, Center, and Spread of the Sampling Distribution of \( \bar{x} \)**

D4. Why is it the case that the sampling distribution of the mean for samples of size 1 is identical to the population distribution?

D5. The scatterplot in Display 7.26 shows the standard error, \( SE \), plotted against the sample size, \( n \), for the table in Display 7.25 on page 429. Find a transformation that linearizes these points. Use the transformation and the equation of the resulting least squares line to justify the rule \( SE = \sigma / \sqrt{n} \).

![Display 7.26 SE plotted against n.](image)

D6. Justify the comment “Large samples are better, because the sample mean tends to be closer to the population mean.”

**Finding Probabilities Involving Sample Means**

As you will see in the next example, you can solve problems involving the sample mean by using the three properties of the sampling distribution in combination with what you learned about normal distributions in Chapter 2.
Example: Average Number of Children

What is the probability that a random sample of 20 families in the United States will have an average of 1.5 children or fewer?

Solution

By looking at the sampling distribution in Display 7.25 for $n = 20$, you can see that the sampling distribution of the sample mean is approximately normal. The mean of the population is 0.9, and the standard deviation of the population is 1.1. So, for the sampling distribution, we have

$$
\mu_x = \mu = 0.9
$$

$$
\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{1.1}{\sqrt{20}} \approx 0.25
$$

Display 7.27 shows a sketch of this situation.

The $z$-score for the value 1.5 is

$$
z = \frac{x - \text{mean}}{\text{standard deviation}} = \frac{x - \mu_x}{\sigma_x} = \frac{1.5 - 0.9}{0.25} \approx 2.4
$$

You can use a table or a calculator to find that the area under a standard normal curve to the left of 1.5 (a $z$-score of 2.4) is about 0.9918, which is the probability that the sample mean will fall below 1.5. In a random sample of 20 families, it is almost certain that the average number of children per family will be less than 1.5.
Example: Reasonably Likely Averages

What average numbers of children are reasonably likely in a random sample of 20 families?

Solution

From the preceding example, the distribution of the mean number of children is approximately normally distributed with mean 0.9 and standard error 0.25. Reasonably likely outcomes are those within approximately two standard errors of the mean. This interval is

$$0.9 \pm 1.96(0.25)$$

which gives an average number of children between 0.41 and 1.39.

Finding Probabilities Involving Sample Means

D7. What is the probability that a random sample of nine U.S. families will have an average of 1.5 children or fewer? Can this problem be done like the example for 20 families? Why or why not?

Using the Properties of the Sampling Distribution of the Mean

At this stage, you must have a few questions about how to use the properties in the box on page 430.

1. *When can I use the property that the mean of the sampling distribution of the mean is equal to the mean of the population, $\mu_{\bar{x}} = \mu$?*
   
   Good news. In random sampling, it is always true that $\mu_{\bar{x}} = \mu$. The shape of the population doesn't matter, nor does how large the sample is, how large the population is, or whether you sample with replacement or without replacement.

2. *When can I use the property that the standard error of the sampling distribution of the mean, $\sigma_{\bar{x}}$, is equal to $\sigma/\sqrt{n}$?*
   
   Good news here too, mostly. You can use this formula with a population of any shape and with any sample size as long as you randomly sample with replacement or you randomly sample without replacement and the sample size is less than 10% of the population size. Almost always, a random sample is taken without replacement but the population size is more than 10 times the sample size. So, almost always, you can use this formula. However, in the rare instance when you are taking a random sample without replacement from a small population, you should use the formula given in E30.

3. *To compute probabilities, I find a z-score and then use the table of standard normal probabilities on page 824 to get the probability. This doesn't work unless the sampling distribution of the mean is approximately normal. When can I assume that the sampling distribution is approximately normal?*

   Here's where the trouble starts. As you saw in Display 7.25, for example, when the population is skewed, the sampling distribution also is skewed for
small sample sizes. With a large enough sample, the sampling distribution is 
approximately normal. The problem is that there is no hard-and-fast rule to 
follow in knowing what sample size is "large enough." You will learn some 
guidelines in Section 9.3.

For now, if you are told that the population is approximately normally 
distributed (or mound-shaped or bell-shaped), you can assume that all 
sampling distributions of the mean are approximately normal too, no matter 
what the sample size. If you are told that the sample size is very large, it's 
safe to assume that the sampling distribution of the mean is approximately 
normal. If the population looks like the one in Display 7.25, you know the 
sampling distribution is approximately normal for samples of size 40 or 
larger, and a normal approximation wouldn't be too bad for samples of size 
20 or larger.

4. Isn't the size of the population really important? Surely a random sample of 
200 households would provide more information about a city of size 20,000 
households than about a city with 2 million households.

Not really. As long as the sample was randomly selected and as long as the 
population is much larger than the sample, it doesn't matter how large the 
population is. Perhaps an extreme example will help. Suppose you want to 
know the percentage of all Skittles that are yellow. You start sampling Skittles. 
After checking 100,000 pieces, you find that about 20% are yellow. With such 
a large random sample, you should have a lot of faith that this estimate is 
close to the real population percentage. You shouldn't have less faith in this 
estimate if you know that 100 million Skittles have been produced than if you 
know 10 million pieces have been produced.

DISCUSSION

Using the Properties of the Sampling Distribution 
of the Mean

D8. If you select 100 households at random, would the standard error of the 
sampling distribution of the mean number of children be larger if the 
population size, \( N \), is 1,000 or if it is 10,000?

D9. Consider the sampling distribution of the mean of a random sample of size 
\( n \) taken from a population of size \( N \) with mean \( \mu \) and standard deviation \( \sigma \).
   a. For a fixed \( N \), how does the mean of the sampling distribution change as 
      \( n \) increases?
   b. For a fixed \( N \), how does the standard error change as \( n \) increases?
   c. If \( N = n \), what are the mean and standard error of the sampling 
      distribution?

Finding Probabilities Involving Sample Totals

Sometimes situations are stated in terms of the total number in the sample rather 
than the average number: “What is the probability that there are 30 or fewer
children in a random sample of 20 families in the United States?” You have the choice of two equivalent ways of doing this problem.

- **Method I:** Find the equivalent average number of children, \( \bar{x} \), by dividing the total number of children, 30, by the sample size, 20:
  \[
  \bar{x} = \frac{30}{20} = 1.5
  \]

  Then you can use the same formulas and procedure as in the previous examples.

- **Method II:** Convert the formulas from the previous examples into equivalent formulas for the sum, and then proceed as described here.

  The sum (or total) of a sample is equal to the sample size times the sample mean:
  \[
  \sum x = n \bar{x}
  \]

  You can get the mean and the standard error of the sampling distribution of the sum from those of the sampling distribution of the mean simply by multiplying by \( n \).

**Properties of the Sampling Distribution of the Sum of a Sample**

If a random sample of size \( n \) is selected from a distribution with mean \( \mu \) and standard deviation \( \sigma \), then

- the mean of the sampling distribution of the sum is
  \[
  \mu_{\text{sum}} = n\mu
  \]

- the standard error of the sampling distribution of the sum is
  \[
  \sigma_{\text{sum}} = n\sigma_{\bar{x}}
  \]
  \[
  = n \frac{\sigma}{\sqrt{n}}
  \]
  \[
  = \sqrt{n} \cdot \sigma
  \]

- the shape of the sampling distribution will be approximately normal if the population is approximately normally distributed; and for other populations the sampling distribution will become more normal as \( n \) increases

**Example: The Probability of 30 or Fewer Children**

What is the probability that a random sample of 20 families in the United States will have a total of 30 or fewer children?
Solution

The sampling distribution of the sum is approximately normal because it has the same shape as the sampling distribution of the mean. It has mean \( \mu_{\text{sum}} = n \mu = 20(0.9) \), or 18, and standard error \( \sigma_{\text{sum}} = \sqrt{n} \cdot \sigma = \sqrt{20}(1.1) \), or approximately 5. The z-score for a total of 30 children is

\[
z = \frac{\text{sample sum} - \mu_{\text{sum}}}{\sigma_{\text{sum}}} = \frac{30 - 20(0.9)}{\sqrt{20}(1.1)} \approx 2.4
\]

Using a calculator or Table A on page 824, the probability that \( z \) is less than 2.40 is about 0.9918.

Example: Reasonably Likely Totals

In a random sample of 20 families, what total numbers of children are reasonably likely?

Solution

From the preceding example, the sampling distribution of the total number of children is approximately normal with mean 18 and standard error 5. The reasonably likely outcomes are those within approximately two standard errors of the mean. This is the interval

\[
18 \pm 1.96(5)
\]

or a total number of children between 8.2 and 27.8.

Finding Probabilities Involving Sample Totals

D10. How do the shapes of the sampling distributions of the sample mean and of the sample total compare? Show how to convert the plots for the sample mean in Display 7.25 on page 429 to plots for the sample total.

Summary 7.2: Sampling Distribution of the Sample Mean

Because averages are so important, it seems fortunate that the mean and standard error of the sampling distribution of \( \bar{x} \) have such simple formulas. It seems doubly fortunate that, for large sample sizes, the shape is approximately normal. This allows you to use z-scores and the standard normal distribution to determine whether a given sample mean is reasonably likely or a rare event. But fortune has nothing to do with it! It's precisely because these formulas are so simple that the mean and standard deviation are so important.

Here is a summary of the characteristics of the sampling distribution of \( \bar{x} \). They apply only if the samples are selected at random.
• The mean of the sampling distribution equals the mean of the population, or $\mu_x = \mu$, and this fact does not depend on the shape of the population or on the sample size.

• The standard error of the sampling distribution equals the population standard deviation divided by the square root of the sample size, or $\sigma_x = \sigma / \sqrt{n}$, and this fact does not depend on the shape of the population or on the sample size as long as you sample with replacement or your sample size is no more than 10% of the size of the population.

• If the population is normally distributed, the sampling distribution is normally distributed for all sample sizes $n$. If the population is not normally distributed, then as the sample size increases, the sampling distribution will become more normal. This is called the Central Limit Theorem.

• If the population is at least ten times the size of the sample, whether you sample with or without replacement is of little consequence.

• The population size does not have much effect on the analysis unless the sample size is greater than 10% of the population size.

If you have a situation involving the sum or total in a sample, you can find a sample mean by dividing the total by $n$. Alternatively, the sampling distribution of the sample sum has mean

$$\mu_{\text{sum}} = n\mu$$

and standard error

$$\sigma_{\text{sum}} = \sqrt{n} \cdot \sigma$$

What rule can you use to decide whether the shape of your sampling distribution is approximately normal? Unfortunately, there isn’t one—the required size of $n$ depends on how close the population itself is to normal and how accurate you want your approximations to be. You will see more on this topic later.

**Practice**

**Shape, Center, and Spread of the Sampling Distribution of $\bar{x}$**

P7. The distribution of the population of the number of motor vehicles per household and sampling distributions of the mean for samples of size 4 and 10 are shown in Display 7.28.

a. Which distribution is which? Make a rough estimate of the mean and standard deviation of each distribution.
b. The mean of the population is about 1.7. Theoretically, what is the mean of the sampling distribution for samples of size 4? Of size 10? Are these computed means consistent with your estimates of the means in the histograms?

c. The standard deviation of the population is about 1. Theoretically, what is the SE of the sampling distribution for samples of size 4? Of size 10? Are these computed SEs consistent with your estimates of the means in the histograms?

d. Compare the shapes of the three distributions. Are the shapes consistent with the Central Limit Theorem?

P8. From 1910 through 1919, the single-season batting averages of individual Major League Baseball players had a distribution that was approximately normal, with mean .266 and standard deviation .037. Suppose you construct the sampling distribution of the mean batting average for random samples of 15 players. What are the shape, mean, and standard error of this distribution? [Source: Stephen Jay Gould, *The Spread of Excellence from Plato to Darwin* (New York: Harmony Books, 1996).]

Finding Probabilities Involving Sample Means

P9. Refer to Display 7.24 on page 428. Suppose a television network selects a random sample of 1000 families in the United States for a survey on TV viewing habits.

a. Describe the distribution of the possible values of the average number of children per family.

b. What average numbers of children are reasonably likely?

c. What is the probability that the average number of children per family will be 0.8 or less?

P10. Last January 1, Jenny thought about buying individual stocks. Over the next year, the mean of the percentage increases in individual stock prices is 6.5% and the standard deviation of these percentage increases is 12.8%. The distribution of price increases is approximately normal.

a. If Jenny had picked one stock at random, what is the probability that it would have gone down in price?

b. If Jenny had picked four stocks at random, what is the probability that their mean percentage increase would be negative?

c. If Jenny had picked eight stocks at random, what is the probability that their mean percentage increase would be between 8% and 10%?

d. If Jenny had picked eight stocks at random, what mean percentage changes in price are reasonably likely? (In other words, find the interval that contains the middle 95% of possible mean percentage increases in value.)

Using the Properties of the Sampling Distribution of the Mean

P11. In this problem, you will review sampling with and without replacement and the rules of probability that you learned in Chapter 5. Suppose you take a random sample of two families from the population in Display 7.29.
Number of Children | Number of Families
--- | ---
0 | 53
1 | 20
2 | 17
3 | 7
4 (or more) | 3
Total | 100

Display 7.29  Population of families and their numbers of children.

a. If you sample with replacement, what is the probability that both families have exactly one child?
b. If you sample without replacement, what is the probability that both families have exactly one child?
c. If you sample with replacement, what is the probability that neither family has exactly one child?
d. If you sample without replacement, what is the probability that neither family has exactly one child?
e. If you sample with replacement, what is the probability that at least one of the families has exactly one child?
f. If you sample without replacement, what is the probability that at least one of the families has exactly one child?

P12. A typical political opinion poll in the United States questions about 1500 people. How would the analysis change if this were a random sample of 1500 people from Arizona rather than from the entire United States?

P13. The distribution of the number of motor vehicles per household in the United States is roughly symmetric, with mean 1.7 and standard deviation 1.0.
a. If you pick 15 households at random, what is the probability that they have at least 30 motor vehicles among them?
b. If you pick 20 households at random, what is the probability that they have between 25 and 30 motor vehicles among them?

P14. Refer to Display 7.24 on page 428. Suppose a television network selects a random sample of 1000 families in the United States for a survey on TV viewing habits.
a. Do you think it is reasonably likely that a sample of 1000 households will produce at least 1000 children? Explain your reasoning.
b. Suppose the network changes to a random sample of 1200 households. Does this dramatically improve the chances of seeing at least 1000 children in the sampled households? Explain.

P15. Suppose the network in P14 repeatedly uses one of these four sample sizes and computes the mean number of children per family. What interval should contain the middle 95% of the sample means? That is, for each sample size, determine what sample means are reasonably likely.
a. 25  
b. 100  
c. 1000  
d. 4000 (This is the approximate sample size for the Nielsen television ratings.)
E15. Consider the skewed population shown in Display 7.30. Display 7.31 shows the three simulated sampling distributions of the sample mean for samples of size 2, 4, and 25.

**Display 7.30** A skewed population.

a. Match the theoretical summary information with the correct simulated sampling distribution (A, B, or C) and the correct sample size (2, 4, or 25).
   I. mean 2.50; standard error 0.48
   II. mean 2.50; standard error 1.20
   III. mean 2.50; standard error 1.70

b. Does the rule for computing the standard error of the mean from the standard deviation of the population appear to hold in all three situations?

c. Does the sampling distribution appear to be approximately normal in all cases? If not, explain how the given shape came about.

d. For which sample sizes would it be reasonable to use the rule stating that about 95% of all sample means lie within approximately two standard errors of the population mean?

**Display 7.31** Three sampling distributions of the sample mean (for \( n = 2, 4, \) and 25, not necessarily in that order).
E16. Consider the M-shaped population in Display 7.32. Display 7.33 shows three simulated sampling distributions of the sample mean for samples of size 2, 4, and 25.

Display 7.32 An M-shaped population.

a. Match the theoretical summary information with the correct simulated sampling distribution (A, B, or C) and the correct sample size (2, 4, or 25).
   I. mean 4.50; standard error 1.75
   II. mean 4.50; standard error 0.70
   III. mean 4.50; standard error 2.47

b. Does the rule for computing the standard error of the mean from the standard deviation of the population appear to hold in all three situations?

c. Does the sampling distribution appear to be approximately normal in all cases? If not, explain how the given shape came about.

d. For which sample sizes would it be reasonable to use the rule stating that about 95% of all sample means lie within approximately two standard errors of the population mean?

Display 7.33 Three sampling distributions (for \( n = 2, 4, \) and 25, not necessarily in that order).
E17. Explain how to use a list of random digits to get a random sample of size 5 from the population in E15. (You will have to estimate the percentages in the population from the histogram.)

E18. Explain how to use a list of random digits to get a random sample of size 5 from the population in E16. (You will have to estimate the percentages in the population from the histogram.)

E19. Suppose police records in a small city show that the number of automobile accidents per day for 1,045 days has the frequency distribution shown in Display 7.34. The relative frequencies, in order, are 0.36, 0.37, 0.17, 0.09, and 0.01.

Display 7.34 Frequency distribution of the number of automobile accidents per day.

a. Would it be a rare event to see two accidents per day in this city?

b. Which histogram is for samples of 4 days? Which is for samples of 8 days?

c. Would it be a rare event to see an average of 1.75 accidents per day over a period of 4 consecutive days in this city?

d. Would it be a rare event to see an average of 1.75 accidents per day over an 8-day period in this city?

e. When you use the distributions in Display 7.35 to answer parts c and d, what assumptions are you making? Do they seem reasonable for this situation?

Display 7.35 Simulated sampling distributions (for \(n = 4\) and \(n = 8\), not necessarily in that order).

E20. The distribution of grades on the very first AP Statistics Exam are given in Display 7.36.

Table: Display 7.36 Distribution of exam grades.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>13.5</td>
</tr>
<tr>
<td>4</td>
<td>21.3</td>
</tr>
<tr>
<td>3</td>
<td>24.6</td>
</tr>
<tr>
<td>2</td>
<td>18.7</td>
</tr>
<tr>
<td>1</td>
<td>21.9</td>
</tr>
</tbody>
</table>

a. Construct a plot of the distribution of the population of scores.

b. The plots in Display 7.37 show simulated sampling distributions of the mean for random samples of size 1, 5, and 25. Match each histogram to its sample size.
7.2 Sampling Distribution of the Sample Mean

A. 1.0 1.4 1.8 2.2 2.6 3.0 3.4 3.8
   Sample Mean
   0 20 40 60
   Frequency

B. 1.0 2.0 3.0
   Sample Mean
   4.0 5.0 10
   0 20 30 40 50
   Frequency

C. 1 1.5 2 2.5
   Sample Mean
   3 3.5 4 4.5
   10 20 30
   40 50
   Frequency

Display 7.37  Simulated sampling distributions
(for n = 1, 5, and 25, not necessarily in that order).

c. A teacher reports that her class averaged 3.6 on this exam. Do you think she has a class of 5 students or a class of 25 students? Explain your reasoning.

E21. A particular college entrance examination has scores that are approximately normally distributed with mean 500 and standard deviation 100.

a. If you pick 40 scores at random, what is the probability that their mean will be within 10 points of the population mean of 500?

b. How large a random sample of scores would you need in order to be 95% sure that the sample mean would be within 10 points of the population mean of 500?

E22. The process for manufacturing a ball bearing results in weights that have an approximately normal distribution with mean 0.15 g and standard deviation 0.003 g.

a. If you select one ball bearing at random, what is the probability that it weighs less than 0.148 g?

b. If you select four ball bearings at random, what is the probability that their mean weight is less than 0.148 g?

c. If you select ten ball bearings at random, what is the probability that their mean weight is less than 0.148 g?

E23. A particular college entrance examination has scores that are approximately normally distributed with mean 500 and standard deviation 100.

a. If you pick 40 scores at random, what is the probability that their mean will be within 10 points of the population mean of 500?

b. How large a random sample of scores would you need in order to be 95% sure that the sample mean would be within 10 points of the population mean of 500?

E24. Last January 1, Jenny thought about buying individual stocks. Over the next year, the mean of the percentage increases in individual stock prices is 6.5% and the standard deviation of these percentage increases is 12.8%. The distribution of price increases is approximately normal.

a. If Jenny picked ten stocks at random, what is the probability that her mean percentage increase in price is more than 7%?

b. What is the smallest number of stocks Jenny should pick at random so that the probability is 95% or more that her mean percentage increase in price will be more than 5%?

E25. An emergency room in Oxford, England, sees a mean of 67.4 children ages 7–15 per summer weekend, with standard deviation 10.4. The numbers of children are approximately normally distributed.


a. How many children in all would you expect to be seen in the emergency room
over two randomly selected summer weekends?
b. Find the probability that the emergency room will see a total of 73 or fewer children over two randomly selected summer weekends.
c. Over the two summer weekends when Harry Potter books were released, a total of 73 children were seen in the emergency room. What can you conclude?

E26. In a recent year, 16,597,552 people were admitted to riverboat casinos in Illinois. The mean casino win (customer loss) per admission was $103.02. Assume that the standard deviation might be about $100. [Source: Illinois Statistical Abstract 2004, Table 27-19; www.igpa.uillinois.edu.]
   a. What were the total winnings for the riverboat casinos that year?
   b. If you randomly select 100 people who were admitted, what is the probability that the total winnings for the casinos from these people was over $11,000?
   c. If you randomly select 100 people who were admitted, what is the probability that the total winnings for the casinos from these people was negative (i.e., the group as a whole won money)?

E27. The advertising agency for a new family-oriented theme park has run a promotion that promises 100 local people free admission for a year. In preparation for a public drawing, the address of each household in the community is placed in a bin. Winning addresses will be drawn, and each member of the household at those addresses will get free admission for a year. The director of the advertising agency suddenly realizes a flaw in his scheme: He doesn’t know how many addresses to draw in order to end up with 100 people. (The number of people in each household won’t be known until the winners are notified.)
   His budget allows for 100 free admissions and he doesn’t want to go over this, as it will cut into his profit. He finds from census data that the distribution of the number of people per household in the community is mound-shaped, with a mean of 4.3 and a standard deviation of 1.4.
   a. What is the smallest number of addresses that should be drawn in order to have a 98% chance that there will be at least 100 people total in the households? (You can use your calculator to estimate the solution to the resulting equation to the nearest whole number of names.)
   b. The cost of one free admission for a year is $250. If the director draws the number of addresses you recommended in part a, how much will he have to pay because of his poor planning?

E28. Display 7.38 gives the distribution of the number of households in the United States that own various numbers of color televisions. Suppose a government agency wants to check 1000 color televisions to assess whether televisions in homes have a tendency to overheat. How large a random sample of households should the agency take to have a 95% chance of getting 1000 color televisions or more among the households in its sample?

<table>
<thead>
<tr>
<th>Number of Color Televisions</th>
<th>Percent of Households</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.2</td>
</tr>
<tr>
<td>1</td>
<td>27.4</td>
</tr>
<tr>
<td>2</td>
<td>35.9</td>
</tr>
<tr>
<td>3</td>
<td>21.8</td>
</tr>
<tr>
<td>4</td>
<td>9.5</td>
</tr>
<tr>
<td>5</td>
<td>4.2</td>
</tr>
</tbody>
</table>

Display 7.38 Number of color televisions per household. [Source: U.S. Census Bureau, Statistical Abstract of the United States, 2006, Table 963.]

E29. As the sample size $n$ increases, describe what happens to the mean and standard error of the sampling distribution of
   a. the sample mean   b. the sample total

E30. If you have a small population and you sample without replacement, the formula for the standard error of the sampling distribution of the mean must be adjusted:

$$SE = \sigma_x = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N - n}{N - 1}}$$
7.2 Sampling Distribution of the Sample Mean

a. Does multiplying by \( \frac{N-n}{N-1} \) increase or decrease the SE? Why is that a sensible thing to do?

In the rest of this exercise, you will see how to use this formula.

You are allowed to drop your two lowest grades on the six tests in your statistics class. There will be no make-up tests. If you showed up for all six tests, your scores would have been 55, 65, 75, 80, 90, 95.

b. What are the mean, \( \mu \), and standard deviation, \( \sigma \), of this population of test scores?

c. Suppose you pick two test days at random to miss class. You get a 0 on these two tests, and these two 0's will be the grades you drop. Construct the exact sampling distribution of your mean test score for the remaining four tests.

d. Verify that the mean of the sampling distribution is equal to the mean of the population of test scores, that is, \( \mu = \mu_x \).

e. The basic formula for the standard deviation of a population is

\[
\sigma = \sqrt{\frac{\sum (x - \mu)^2}{n}}
\]

Use this formula to compute the standard error of the sampling distribution from part c. Note that there are 15 values in this sampling distribution, so \( n = 15 \).

f. Compute the standard error (SE) of the sampling distribution using the two short-cut formulas:

\[
\sigma_x = \frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \sigma_x = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}
\]

g. Which formula in part f gives the correct SE?

Use this information for E31–E34: In Section 6.1, you learned rules for finding the mean and variance when you add or subtract two random variables in a probability distribution (see the box on page 372). The same rules apply when working with sampling distributions. Here the rules are stated in terms of sampling distributions.

---

**Properties of the Sampling Distribution of the Sum and Difference**

Suppose two values are taken randomly from two populations with means \( \mu_1 \) and \( \mu_2 \) and variances \( \sigma_1^2 \) and \( \sigma_2^2 \), respectively. Then the sampling distribution of the sum of the two values has mean

\[
\mu_{\text{sum}} = \mu_1 + \mu_2
\]

If the two values were selected independently, the variance of the sum is

\[
\sigma_{\text{sum}}^2 = \sigma_1^2 + \sigma_2^2
\]

The sampling distribution of the difference of the two values has mean

\[
\mu_{\text{difference}} = \mu_1 - \mu_2
\]

If the two values were selected independently, the variance of the difference is

\[
\sigma_{\text{difference}}^2 = \sigma_1^2 + \sigma_2^2
\]

The shapes of the sampling distributions of the sum and the difference depend on the shapes of the two original populations. If both populations are normally distributed, so are the sampling distributions of the sum and the difference.

---

E31. In 2006, SAT critical reading scores had a mean of 503 and a standard deviation of 113. The SAT math scores had a mean of 518 and a standard deviation of 115. Each distribution was approximately normal.

**Source:** 2006 College Bound Seniors: Total Group Profile Report, www.collegeboard.com

a. Suppose one SAT critical reading score is selected at random and one SAT math score is independently selected at random and the scores are added. Find the mean and standard error of the sampling distribution of this sum.

b. Find the probability that the sum of the scores from part a is less than 800.

c. What total SAT scores are reasonably likely?
d. What is the probability that the critical reading score is at least 100 points higher than the math score?

e. Suppose one student is selected at random and his or her SAT critical reading score and math score are added. If you can, describe the mean and standard error of the sampling distribution of this sum.

E32. The height of males ages 20–29 in the United States is approximately normally distributed with mean 69.3 inches and standard deviation 2.92. The height of females ages 20–29 in the United States is approximately normally distributed with mean 64.1 inches and standard deviation 2.75. [Source: U.S. Census Bureau, *Statistical Abstract of the United States*, 2006, Table 196.]

a. Suppose one male from this age group is selected at random and one female is independently selected at random and their heights added. Find the mean and standard error of the sampling distribution of this sum.

b. Find the probability that the sum of the heights is less than 125 inches.

c. What total heights are reasonably likely?

d. What is the probability that the male is at least 2 inches taller than the female?

e. Suppose that one male is selected at random and his height is added to that of his closest female relative aged 20 to 29. If you can, describe the mean and standard error of the sampling distribution of this sum.

E33. In the case where $n$ values are taken at random and independently from the same population with mean $\mu$ and standard deviation $\sigma$, the sampling distribution of their sum has mean $\mu + \mu + \cdots + \mu$, or $n\mu$, and variance $\sigma^2 + \sigma^2 + \cdots + \sigma^2$, or $n\sigma^2$.

a. If you roll a die three times, what is the mean of the sampling distribution of the sum? The variance?

b. If you roll a die seven times, what is the mean of the sampling distribution of the sum? The variance?

c. In 2006, SAT critical reading scores had a mean 503 and standard deviation 113. If 20 scores are selected at random, what are the mean and variance of their sum? What is the probability that the sum is less than 10,000?

E34. F. Paca weighed rams in Lima, Peru, and found that the distribution was approximately normal with mean 75.4 kg and standard deviation 7.38 kg. [Source: F. Paca (New Mexico State University, 1977).]

a. Suppose two rams are selected at random. Describe the sampling distribution of the sum of their weights.

b. What is the probability that the sum of their weights is less than 145 kg?

c. What is the probability that the sum of the weights of ten randomly selected rams is more than 750 kg? (Use the rules in E33.)

d. What is the probability that the first ram selected is more than 5 kg heavier than the second ram selected?

### 7.3 Sampling Distribution of the Sample Proportion

You now will move from studying the behavior of sample means to studying the behavior of sample proportions (the proportion of “successes” in the sample). You will see that the properties of sample proportions parallel those of sample means.

### Sampling Distribution of the Number of Successes

Before moving on to sample proportions, you will review what you have learned so far about the number of successes you can expect in a binomial
(success/failure) situation. The box and example summarize relevant information from Section 6.2.

### Properties of the Sampling Distribution of the Number of Successes

If a random sample of size \( n \) is selected from a population with proportion of successes \( p \), then the sampling distribution of the number of successes \( X \)

- has mean \( \mu_X = np \)
- has standard error \( \sigma_X = \sqrt{np(1-p)} \)
- will be approximately normal as long as \( n \) is large enough

As a guideline, if both \( np \) and \( n(1-p) \) are at least 10, then using the normal distribution as an approximation to the shape of the sampling distribution will give reasonably accurate results.

### Example: Probability of 30 or More Mississippians Wearing Seat Belts

The use of seat belts continues to rise in the United States, with overall seat belt usage of 82%. Mississippi lags behind the rest of the nation—only about 60% wear seat belts. [Source: National Highway Traffic Safety Administration, November 2005, www.nhtsa.dot.gov.] Suppose you take a random sample of 40 Mississippians. How many do you expect will wear seat belts? What is the probability that 30 or more of the people in your sample wear seat belts?

**Solution**

You expect that 60% of the 40, or 24, Mississippians will be wearing seat belts.

To estimate the probability that 30 or more wear seat belts, you will use the normal approximation to the binomial distribution. However, first you must check the guideline that both \( np \) and \( n(1-p) \) are at least 10. This ensures that the binomial distribution is sufficiently mound-shaped for the normal distribution to be a reasonable model.

Because 60% of the population use seat belts, both \( np \) and \( n(1-p) \) are 10 or greater:

\[
np = 40(0.6) = 24 \\
n(1-p) = 40(1-0.6) = 16
\]

Thus, the sampling distribution of the number of successes is approximately normal, with mean and standard error

\[
\mu_X = np = 40(0.6) = 24 \\
\sigma_X = \sqrt{np(1-p)} = \sqrt{40(0.6)(1-0.6)} \approx 3.098
\]

This distribution is shown in Display 7.39.
Display 7.39  Exact sampling distribution and normal approximation of the number of successes when $p = 0.6$ and $n = 40$.

The $z$-score for 30 successes is

$$z = \frac{x - \mu_x}{\sigma_{\text{sum}}} = \frac{x - np}{\sqrt{np(1-p)}} \approx \frac{30 - 24}{3.098} = 1.937$$

The proportion below this value of $z$ is 0.9736, so the probability of getting 30 or more successes is $1 - 0.9736$, or 0.0264.

**Sampling Distribution of the Sample Proportion**

If you take a random sample of 40 Mississippians, what is the probability that 75% or more wear seat belts? This is the same question answered in the previous example, because 30 out of 40 is 75%. However, the notation (and the computations) will reflect that you now are working with the proportion of successes in the sample rather than the number of successes.

Some notation will make your work more efficient. Suppose you count the number of successes in a random sample of size $n$ from a population in which the true proportion of successes is given by $p$. The symbol for the sample proportion is $\hat{p}$ (read “p-hat”). That is,

$$\hat{p} = \frac{\text{number of “successes”}}{\text{sample size}}$$

For example, suppose your sample of automobile drivers contains 26 who use seat belts. Then $n = 40$, $p = 0.60$, and $\hat{p} = \frac{26}{40} = 0.65$.

In Activity 7.3a, you will construct a simulated sampling distribution of the proportion of drivers wearing seat belts in samples of size 10, 20, and 40.

**ACTIVITY 7.3a**

**Buckle Up!**

**What you’ll need:** a table of random digits

1. Describe how to use a table of random digits to select a random sample of size 10 from a population with 60% “successes.”

2. Use your method to select a random sample of size 10. Count the number of “successes,” and compute the sample proportion, $\hat{p}$. 

(continued)
3. Repeat step 2 until your class has 100 samples of size 10. 
4. Plot your 100 sample proportions on a dot plot. 
5. Calculate the mean and standard error of your simulated sampling distribution of the sample proportion. 
6. Repeat steps 2 through 5 for samples of size 20 and size 40. 

[See Calculator Note 7B to learn how to use a calculator to sample for this activity.]

Display 7.40 shows the exact sampling distributions for samples of size 10, 20, and 40 drawn from a population with 60% “successes.” They should be similar in shape, center, and spread to your dot plots from Activity 7.3a.

Display 7.40  Exact sampling distributions of \( \hat{p} \) for samples of size 10, 20, and 40 when \( p = 0.60 \).

**DISCUSSION**

**Sampling Distribution of the Sample Proportion**

D11. Use Display 7.40 to answer these questions.

a. Which distribution is most normal in shape? Least normal? Does the guideline that both \( np \) and \( n(1 - p) \) are at least 10 work well in predicting approximate normality? (“Approximately normal” means a histogram that is symmetric and mound-shaped and has enough bars that the shape is fairly smooth rather than lumpy.)

b. About 60% of Mississippians wear seat belts. What proportion of Mississippians would you expect to be wearing seat belts in a sample of size 10? Size 20? Size 40?

c. Which distribution in Display 7.40 has the largest spread? The smallest?

d. Would you be most likely to find that 75% or more of Mississippians in the sample wear seat belts if the sample size is 10, 20, or 40?

**Shape, Center, and Spread for Sample Proportions**

Dividing by the sample size, \( n \), to change from the number of successes to the proportion of successes doesn't change the shape of the sampling distribution. Thus, you still can use the guideline that if both \( np \) and \( n(1 - p) \) are at least 10, then the sampling distribution is approximately normal.
Finding the formulas for the mean, \( \mu_p \), and the standard error, \( \sigma_p \), of the sampling distribution of the sample proportion is simple: Just as you divide by \( n \) to convert the number of successes to the proportion of successes, you divide by \( n \) to convert the formulas for the mean, \( \mu_x \), and standard error, \( \sigma_x \), of the number of successes to the equivalent formulas for the proportion of successes:

\[
\mu_p = \frac{\mu_X}{n} = \frac{np}{n} = p
\]

\[
\sigma_p = \frac{\sigma_X}{\sqrt{n}} = \frac{\sqrt{np(1 - p)}}{\sqrt{n}} = \sqrt{\frac{np(1 - p)}{n^2}} = \sqrt{\frac{p(1 - p)}{n}}
\]

These formulas tell you two facts about the sampling distribution of \( \hat{p} \):

- The mean does not change depending on the sample size. No matter how large the sample size, the mean stays at \( p \).
- The spread decreases as the sample size increases.

The properties of the sampling distribution of sample proportions are summarized in this box, followed by an example showing how to use them.

### Properties of the Sampling Distribution of the Sample Proportion

If a random sample of size \( n \) is selected from a population with proportion of successes \( p \), then the sampling distribution of \( \hat{p} \) has these properties:

- The mean of the sampling distribution is equal to the mean of the population, or \( \mu_{\hat{p}} = p \).
- The standard error of the sampling distribution is equal to the standard deviation of the population divided by the square root of the sample size:

\[
\sigma_{\hat{p}} = \frac{\sigma_{\hat{x}}}{\sqrt{n}} = \frac{\sqrt{np(1 - p)}}{\sqrt{n}} = \sqrt{\frac{np(1 - p)}{n}}
\]

- As the sample size gets larger, the shape of the sampling distribution becomes more normal and will be approximately normal if \( n \) is large enough.

As a guideline, if both \( np \) and \( n(1 - p) \) are at least 10, then using the normal distribution as an approximation of the shape of the sampling distribution will give reasonably accurate results.

### Example: Describing the Sampling Distribution of a Sample Proportion

Drivers in the Northeast and Mid-Atlantic states had the highest failure rate, 20%, on the GMAC Insurance National Driver’s Test. (They also were the drivers most likely to speed.) [Source: Insurance Journal, www.insurancejournal.com.] Describe the shape, center, and spread of the sampling distribution of the proportion of drivers who would fail the test in a random sample of 60 drivers from these states. What are the reasonably likely proportions of drivers who would fail the test?
**Solution**

Both \( np = 60(0.2) = 12 \) and \( n(1 - p) = 60(0.8) = 48 \) are at least 10, so the sampling distribution is approximately normal.

The mean, \( \hat{p} \), is 0.2. The SE is

\[
\sigma_{\hat{p}} = \sqrt{\frac{p(1 - p)}{n}} = \sqrt{\frac{0.2(1 - 0.2)}{60}} \approx 0.05
\]

Thus, the reasonably likely proportions are \( 0.2 \pm 1.96(0.05) \), or between about 0.1 and 0.3. The histogram in Display 7.41 shows the exact distribution and normal approximation.

**DISCUSSION**

**Shape, Center, and Spread for Sample Proportions**

D12. Look back at Display 7.40 on page 449.

a. Compute the mean and standard error for each distribution (\( n = 10, 20, \) and 40) using the formulas from this section.

b. Estimate the means and standard errors simply from looking at the histograms. Do the computed means and standard errors in part a match your estimates?

c. The table in Display 7.42 gives the exact sampling distribution for samples of size 10, with \( p = 0.6 \). Compute the mean and SE using the formulas from Section 6.1 given on the next page. Do these formulas give the same mean and SE as the formulas you used in part a?

<table>
<thead>
<tr>
<th>( \hat{p} )</th>
<th>Probability</th>
<th>( \hat{p} )</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.000105</td>
<td>0.6</td>
<td>0.250823</td>
</tr>
<tr>
<td>0.1</td>
<td>0.001573</td>
<td>0.7</td>
<td>0.214991</td>
</tr>
<tr>
<td>0.2</td>
<td>0.010617</td>
<td>0.8</td>
<td>0.120932</td>
</tr>
<tr>
<td>0.3</td>
<td>0.042467</td>
<td>0.9</td>
<td>0.040311</td>
</tr>
<tr>
<td>0.4</td>
<td>0.111477</td>
<td>1.0</td>
<td>0.006047</td>
</tr>
<tr>
<td>0.5</td>
<td>0.200658</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Display 7.42 Exact sampling distribution for samples of size 10, with \( p = 0.6 \).
The expected value (mean) $\mu_X$ and variance $\sigma^2_X$ of a probability distribution listed in a table are

$$E(X) = \mu_X = \sum x_i p_i \quad \text{and} \quad \text{Var}(X) = \sigma^2_X = \sum (x_i - \mu_X)^2 p_i$$

where $p_i$ is the probability that the random variable takes on the specific value $x_i$. The standard deviation is the square root of the variance.

**Finding Probabilities Involving Proportions**

You can now find answers to specific questions by using your knowledge of the properties of the sampling distribution of a proportion.

**Example: Using the Properties of a Sampling Distribution of a Proportion to Find Probabilities**

About 60% of Mississippian use seat belts. Suppose your class conducts a survey of 40 randomly selected Mississippians.

a. What is the chance that 75% or more of those selected wear seat belts?

b. Would it be quite unusual to find that fewer than 25% of the Mississippians selected wear seat belts?

**Solution**

From the developments in this section, you know that $\hat{p}$ has a sampling distribution with mean and standard error

$$\mu_{\hat{p}} = p = 0.6$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1 - p)}{n}} = \sqrt{\frac{0.6(1 - 0.6)}{40}} = 0.0775$$

Furthermore, you can consider the sampling distribution to be approximately normal because $np$ and $n(1 - p)$ are both at least 10:

$$np = 40(0.6) = 24 \quad \text{and} \quad n(1 - p) = 40(1 - 0.6) = 16$$

This distribution is sketched in Display 7.43.

![Display 7.43](image-url) Exact and approximate sampling distribution of $\hat{p}$ for $p = 0.6$ and $n = 40.
a. The z-score for the value $\hat{p} = 0.75$ in this distribution is

$$z = \frac{\hat{p} - \mu_p}{\sigma_{\hat{p}}} = \frac{0.75 - 0.6}{\sqrt{0.0775/40}} \approx 1.935$$

The proportion below this value of $z$ is 0.9735. It follows that the probability of getting 75% or more Mississippians who wear seat belts in a sample of size 40 is only 1 - 0.9735, or 0.0265.

b. The sample proportion, 0.25, is about 4.5 standard deviations below the mean of the sampling distribution, so such a result would be unusual indeed!

**Summary 7.3: Sampling Distribution of the Sample Proportion**

Sample proportions, like sample means, are among the most common summary statistics in practical use, and knowledge of their behavior is fundamental to the study of statistics.

The formulas in this section about the proportion of successes are derived from those in Section 6.2 about the number of successes. Either set of formulas can be used in ways parallel to those in Section 7.2 for the sample mean. You compute a z-score and use the normal distribution to approximate the probability of getting specified results from a random sample. To do this, you use these facts:

- The samples must be selected randomly.
- The mean $\mu_{\hat{p}}$ of the sampling distribution of $\hat{p}$ is equal to the proportion $p$ of successes in the population, or $\mu_{\hat{p}} = p$.
- The standard deviation $\sigma_{\hat{p}}$ of the sampling distribution of $\hat{p}$, also called the standard error of a sample proportion, is given by the formula

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

- Alternatively, the sampling distribution of the number of successes has mean $\mu_X$ and standard error $\sigma_X$, where

$$\mu_X = np \quad \text{and} \quad \sigma_X = \sqrt{np(1-p)}$$

- The shape of the sampling distribution becomes more normal as $n$ increases. You can assume it is normal for computing probabilities as long as $np$ and $n(1-p)$ are both at least 10.

**Practice**

**Sampling Distribution of the Number of Successes**

P16. A survey of hundreds of thousands of college freshmen found that 63% believe “dissent is a critical component of the political process.”

[Source: Higher Education Research Institute, UCLA, The American Freshman, National Norms for Fall 2005.]

Suppose you take a random sample of 100 of the freshmen surveyed. What is the probability that you will find that between 56 and 70 of the freshmen in your sample believe this?
P17. The GMAC Insurance National Driver’s Test found that about 10% of drivers ages 16 to 65 would fail a written driver’s test if they had to take one today. [Source: Insurance Journal, 2005, www.insurancejournal.com.] The sampling distributions in Display 7.44 give the proportion of drivers who would fail for samples of size 10, 20, 40, and 100 taken from a population like this one, where \( p = 0.10 \).

Display 7.44  Exact sampling distributions of \( \hat{p} \) for samples of size 10, 20, 40, and 100 when \( p = 0.10 \).

a. Which distribution is most normal in shape? Least normal? Does the guideline that both \( np \) and \( n(1 - p) \) should be at least 10 work well in predicting approximate normality?

b. What proportion of drivers would you expect to fail a driving test in a sample of size 10? Size 20? Size 40? Size 100?

c. Which distribution has the largest spread? The smallest?

d. Would you be most likely to find that 20% or more of a sample of drivers fail a driving test if the sample size is 10, 20, 40, or 100?

P18. Refer to the sampling distributions in P17.

a. Compute the mean and standard error for each distribution (\( n = 10, 20, 40, \) and \( 100 \)) using the formulas in this section.

b. Estimate the means and standard errors simply from looking at the histograms. Do the computed means and standard errors in part a match your estimates?

c. The table in Display 7.45 gives the exact sampling distribution for samples of size 10, with \( p = 0.1 \). Compute the mean and \( SE \) using the formulas given in D12 on page 452. Do these formulas give the same mean and \( SE \) as the formulas you used in part a?

<table>
<thead>
<tr>
<th>( \hat{p} )</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.348678</td>
</tr>
<tr>
<td>0.1</td>
<td>0.387420</td>
</tr>
<tr>
<td>0.2</td>
<td>0.193710</td>
</tr>
<tr>
<td>0.3</td>
<td>0.057396</td>
</tr>
<tr>
<td>0.4</td>
<td>0.011160</td>
</tr>
<tr>
<td>0.5</td>
<td>0.001488</td>
</tr>
<tr>
<td>0.6</td>
<td>0.000138</td>
</tr>
<tr>
<td>0.7</td>
<td>0.000098</td>
</tr>
<tr>
<td>0.8</td>
<td>0.000000</td>
</tr>
<tr>
<td>0.9</td>
<td>0.000000</td>
</tr>
<tr>
<td>1.0</td>
<td>0.000000</td>
</tr>
</tbody>
</table>

Display 7.45  Exact sampling distribution for \( n = 10 \) and \( p = 0.1 \).

P19. In the 2000 U.S. Census, 53% of the population over age 30 were women. [Source: U.S. Census Bureau, www.census.gov.]

a. Describe the shape, mean, and standard error of the sampling distribution of \( \hat{p} \) for random samples of size 100 taken from this population. Make an accurate sketch, with a scale on the horizontal axis of this distribution.
b. To be a member of the U.S. Senate, you must be at least 30 years old. In 2000, 9 of the 100 members of the U.S. Senate were women. Is this a reasonably likely event if gender plays no role in whether a person becomes a U.S. Senator?

Members of the United States Congress, 2005

Finding Probabilities Involving Proportions

P20. In 2005, about 31% of the citizens of the United States were Republicans. [Source: Harris poll, March 9, 2005, www.harrisinteractive.com.]

a. If 435 citizens are selected at random, what is the probability of getting 53% or more who are Republicans?

b. In 2006, the U.S. House of Representatives had 435 members, of whom 231 were Republicans. Is this percentage higher than could reasonably occur by chance? If so, what are some possible reasons for this unusually large sample percentage? [Source: clerk.house.gov.]

P21. About 60% of Mississippians wear seat belts. What proportion of seat belt users would be reasonably likely to occur in a random sample

a. of 40 drivers?

b. of 100 drivers?

c. of 400 drivers?

Exercises

E35. The ethnicity of about 92% of the population of China is Han Chinese. Suppose you take a random sample of 1000 Chinese. [Source: CIA World Factbook.]

a. Make an accurate sketch, with a scale on the horizontal axis, of the sampling distribution of the proportion of Han Chinese in your sample.

b. Make an accurate sketch, with a scale on the horizontal axis, of the sampling distribution of the number of Han Chinese in your sample.

c. What is the probability of getting 90% or fewer Han Chinese in your sample?

d. What is the probability of getting 925 or more Han Chinese?

e. What numbers of Han Chinese would be rare events? What proportions?


a. Make an accurate sketch of the sampling distribution of the proportion of people in your sample who have one of these surnames.

b. Make an accurate sketch of the sampling distribution of the number of people in your sample who have one of these surnames.

c. What is the probability of getting 20% or fewer with one of these surnames in your sample?

d. What is the probability of getting 105 or more people with one of these surnames?

e. What numbers of people with one of these surnames would be rare events? What proportions?
E37. Refer to the situation in E35. This time, suppose you take a random sample of 100 Chinese rather than 1000.
   a. Describe how the shape, center, and spread of the sampling distribution of the proportion of Han Chinese in your sample will be different from your sketch in E35, part a.
   b. Describe how the shape, center, and spread of the sampling distribution of the number of Han Chinese in your sample will be different from your sketch in E35, part b.
   c. With a sample of size 100, will the probability of getting 90% or fewer Han Chinese in your sample be larger or smaller than the probability you computed in E35, part c? Explain.

E38. Refer to the situation in E36. This time, suppose you take a random sample of 100 Spanish-surnamed Americans rather than 500.
   a. Describe how the shape, center, and spread of the sampling distribution of the proportion of people in your sample with one of the given surnames will be different from your sketch in E36, part a.
   b. Describe how the shape, center, and spread of the sampling distribution of the number of people with one of the given surnames will be different from your sketch in E36, part b.
   c. With a sample of size 100, will the probability of getting 20% or fewer with one of the given surnames be larger or smaller than the probability you computed in E36, part c? Explain.

E39. In 1991, the median age of residents of the United States was 33.1 years. [Source: U.S. Census Bureau, www.census.gov.]
   a. What is the probability that one person, selected at random, will be under the median age?
   b. In a random sample of 50 people from the United States, what is the probability of getting 10 or fewer under the median age?
   c. Of the 50 Westvaco employees listed in Display 1.1 on page 5, 10 were under the median age. Is this about what you would expect from a random sample of 50 residents of the United States, or should you conclude that this group is special in some way? If you think the group is special, what is special about it?

E40. In fall 2004, 37% of the 38,859 first-year students attending the California State University system needed remedial work in mathematics. [Source: California State University, www.asd.calstate.edu.]
   a. Suppose you select 136 students at random from this population of students. Make an accurate sketch, with a scale on the horizontal axis, of the sampling distribution of the number of students who need remedial work.
   b. What is the probability that 68 or fewer in a random sample of 136 students need remedial work?
   c. Suppose you select 2850 students at random. Make an accurate sketch, with a scale on the horizontal axis, of the sampling distribution of the proportion who need remedial work.
   d. What is the probability of getting 54% or more who need remedial work in a random sample of 2850 students?
   e. Of the 2850 students entering California State University, Northridge, 54% needed remedial work. Is this result about what you would expect from a random sample, or should you conclude that this group is special in some way?
E41. About 60% of married women are employed. If you select 75 married women, what is the probability that between 30 and 40 of them are employed? What assumptions underlie your computation? [Source: U.S. Census Bureau, Statistical Abstract of the United States, 2006.]

E42. Suppose 80% of a certain brand of computer disk contain no bad sectors. If 100 such disks are inspected, what is the approximate chance that 15 or fewer contain bad sectors? What assumptions underlie this approximation?

E43. The histograms in Display 7.46 are sampling distributions of $\hat{p}$ for samples of size 5, 25, and 100, first for a population with $p = 0.2$ and then for a population with $p = 0.4$.
   a. Do the means of the sampling distributions depend on $p$? On $n$?
   b. How do the spreads of the sampling distributions depend on $p$ and $n$?
   c. How do the shapes of the sampling distributions depend on $p$ and $n$?
   d. For which combination(s) of $p$ and $n$ would you be willing to use the rule that roughly 95% of the values lie within two standard errors of the mean?

E44. The guideline states that it is appropriate to use the normal distribution as an approximation for the sampling distribution of a sample proportion if both $np$ and $n(1 - p)$ are greater than or equal to 10. To check this out, generate simulated sampling distributions for values of $p$ equal to 0.90 and 0.98. Use sample sizes of 50, 100, and 500 with each value of $p$. Does the guideline appear to be reasonable?

E45. In this exercise you will learn why the introduction to this section said that the properties of sample proportions parallel the properties of sample means in the previous section. Recall that 60% of Mississippians use seat belts. Imagine the population of Mississippians as consisting of a barrel containing one piece of paper per Mississippian. Those Mississippians who wear seat belts are represented by a piece of paper with the number 1 on it. Those Mississippians who don’t wear seat belts are represented by a piece of paper with the number 0 on it. Display 7.47 (on the next page) shows this population.

Display 7.46 Sampling distributions of $\hat{p}$ for $p = 0.2$ and $p = 0.4$ for samples of size $n = 5$, 25, and 100.
Use Seat Belts? Relative Frequency

<table>
<thead>
<tr>
<th></th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (no)</td>
<td>0.4</td>
</tr>
<tr>
<td>1 (yes)</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Display 7.47 Relative frequency histogram for a population with $p = 0.6$.

a. Instead of taking a random sample of 40 actual Mississippians, you take a random sample of 40 slips of paper out of this barrel. Suppose you get 26 slips of paper with a 1 on them and 14 slips of paper with 0 on them. Compute the sample mean using the formula $\bar{x} = \frac{\sum xf}{n}$. How does this mean compare to the proportion of “successes,” $\hat{p}$, in the sample? Why is that the case?

b. Find the mean of this population (Display 7.47) using the formula $\mu = \sum x \cdot P(x)$. Compute the mean using the formula in the box on page 450. How do the means compare?

c. Describe what you have learned in this exercise about the similarities of a sample proportion and a sample mean.

E46. Refer to E45, where a population with 60% successes is thought of as a population with 60% 1’s and 40% 0’s.

a. Compute the standard error of the population in Display 7.47 using the formula $\sigma = \sqrt{\sum (x - \mu)^2 \cdot P(x)}$.

b. Compute the standard error of the population in Display 7.47 using the formula in the box on page 450. How does this compare with the standard error you computed in part a?

Chapter Summary

In this chapter, you have learned to create a sampling distribution of a summary statistic. To create an approximate sampling distribution using simulation, you first define a process for taking a random sample from the given population. You then take this random sample and compute the summary statistic you are interested in. Finally, you generate a distribution of values of the summary statistic by repeating the process. For some simple situations, you can list all possible samples and get the exact distribution of the summary statistic.

Some summary statistics have easily predictable sampling distributions.

- If a random sample of size $n$ is taken from a population with mean $\mu$ and variance $\sigma^2$, the sampling distribution of the sample mean, $\bar{x}$, has mean and standard error

$$
\mu_{\bar{x}} = \mu \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}
$$

The sampling distribution of the sum of the values in the sample has mean and standard error

$$
\mu_{\text{sum}} = n\mu \quad \text{and} \quad \sigma_{\text{sum}} = \sqrt{n} \cdot \sigma
$$

These formulas are summarized in Display 7.48.
• If a random sample of size \( n \) is taken from a population with percentage of successes \( p \), the sampling distribution of the sample proportion, \( \hat{p} \), has mean and standard error

\[
\mu_{\hat{p}} = p \quad \text{and} \quad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}
\]

The sampling distribution of the number of successes has mean \( np \) and standard error \( \sqrt{np(1-p)} \). These formulas are summarized in Display 7.48.

In the case of means and sums, the sampling distribution becomes more and more normal in shape as the sample size increases. This is called the Central Limit Theorem. If the population is normally distributed to begin with, the sampling distributions of means, sums, and differences are normal for all sample sizes. For proportions, you can use the normal distribution as an approximation to the sampling distribution as long as \( np \) and \( n(1-p) \) are both at least 10.

It might still be unclear to you why sampling distributions are so important. If so, it might help you to review the discussion of the Westvaco case at the beginning of Section 7.1. That sampling distribution helped you decide whether the ages of the workers laid off could reasonably be attributed to the variation that occurs from random sample to random sample, or whether you should investigate if some other mechanism is at work.

There's another reason why sampling distributions are so important. Suppose a pollster takes a random sample of 1500 voters to find out what percentage \( p \) of voters approve of some issue. The proportion \( \hat{p} \) in the sample who approve almost certainly won't be exactly equal to the proportion \( p \) in the population who approve—this is called sampling error. How large is the sampling error likely to be? That's what sampling distributions can tell us. That might seem a bit backward. In this chapter, you always knew \( p \) and then found the reasonably likely values of \( \hat{p} \). But the pollster has \( \hat{p} \) and wants to find plausible values of \( p \). This isn't hard to do now that you know what values of \( \hat{p} \) tend to come from which values of \( p \), but the details will have to wait until Chapter 8.

### Sampling Distribution

<table>
<thead>
<tr>
<th>Center</th>
<th>All Populations</th>
<th>→</th>
<th>Of the Mean</th>
<th>Of the Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \mu_x = \mu )</td>
<td>( n \times \mu )</td>
<td>( \mu_{\text{sum}} = np )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \mu_{\hat{p}} = p )</td>
<td>( n \times p )</td>
<td>( \mu_{\text{sum}} = np )</td>
<td></td>
</tr>
<tr>
<td>Spread</td>
<td>( \sigma_x = \frac{\sigma}{\sqrt{n}} )</td>
<td>( n \times \sigma )</td>
<td>( \sigma_{\text{sum}} = \sqrt{n} \times \sigma )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} )</td>
<td>( n \times )</td>
<td>( \sigma_{\text{sum}} = \sqrt{np(1-p)} )</td>
<td></td>
</tr>
</tbody>
</table>

Display 7.48 | Formulas for the mean and standard error of sampling distributions.
E47. *China Daily*, an English-language newspaper published in the People's Republic of China, reported the Air Pollution Index (API) for 32 major Chinese cities (Display 7.49).

<table>
<thead>
<tr>
<th>City</th>
<th>API</th>
<th>City</th>
<th>API</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beijing</td>
<td>123</td>
<td>Nanchang</td>
<td>56</td>
</tr>
<tr>
<td>Changchun</td>
<td>74</td>
<td>Nanjing</td>
<td>61</td>
</tr>
<tr>
<td>Changsha</td>
<td>77</td>
<td>Nanning</td>
<td>75</td>
</tr>
<tr>
<td>Chengdu</td>
<td>89</td>
<td>Shanghai</td>
<td>75</td>
</tr>
<tr>
<td>Chongqing</td>
<td>77</td>
<td>Shenyang</td>
<td>114</td>
</tr>
<tr>
<td>Fuzhou</td>
<td>86</td>
<td>Shenzhen</td>
<td>42</td>
</tr>
<tr>
<td>Guangzhou</td>
<td>98</td>
<td>Shijiazhuang</td>
<td>218</td>
</tr>
<tr>
<td>Guiyang</td>
<td>85</td>
<td>Taiyuan</td>
<td>346</td>
</tr>
<tr>
<td>Haikou</td>
<td>41</td>
<td>Tianjin</td>
<td>156</td>
</tr>
<tr>
<td>Hangzhou</td>
<td>62</td>
<td>Urumqi</td>
<td>83</td>
</tr>
<tr>
<td>Harbin</td>
<td>138</td>
<td>Wuhan</td>
<td>224</td>
</tr>
<tr>
<td>Hefei</td>
<td>42</td>
<td>Xi'an</td>
<td>221</td>
</tr>
<tr>
<td>Hohhot</td>
<td>164</td>
<td>Xining</td>
<td>500</td>
</tr>
<tr>
<td>Jinan</td>
<td>128</td>
<td>Yantai</td>
<td>76</td>
</tr>
<tr>
<td>Kunming</td>
<td>119</td>
<td>Yinchuan</td>
<td>328</td>
</tr>
<tr>
<td>Lanzhou</td>
<td>500</td>
<td>Zhengzhou</td>
<td>157</td>
</tr>
</tbody>
</table>

Display 7.49  API for Chinese cities. [Source: China Daily, March 13, 1999.]

The mean of this distribution, shown in the dot plot in Display 7.50, is 140.9, with standard deviation 119.4.

Display 7.50  Dot plot of the Air Pollution Index (API) for 32 major Chinese cities.

Suppose China had the resources to monitor air quality in only three of these cities.

a. Describe how to use simulation to construct the sampling distribution of the sample mean, $\bar{x}$, for randomly selected samples of three cities, without replacement. Take one such random sample and compute $\bar{x}$.

b. Find the mean and standard error of the sampling distribution of the sample mean, $\bar{x}$, for samples of size 3.

c. Can the sampling distribution be considered approximately normal?

E48. Refer to the situation in E47 of selecting three Chinese cities at random from the list of 32 cities. Air quality is considered “good” if the API is less than 100.

a. Describe how to use simulation to construct the sampling distribution of the number of cities with good air quality for randomly selected samples of three cities.

b. Find the mean and standard error of the sampling distribution of the number of cities with good air quality for samples of size 3.

c. Can the sampling distribution be considered approximately normal?

E49. About 68% of the people in China live in rural areas. Suppose a random sample of 200 Chinese people is taken.

a. Describe the shape, mean, and standard error of the sampling distribution of $\hat{p}$, the proportion of people in the sample who live in rural areas.

b. What is the probability that 75% or more in the sample live in rural areas?

c. What values of $\hat{p}$ would be rare events?

d. What is the probability that 130 or more in the sample live in rural areas?
E50. The spine widths of the books in a particular library have a distribution that is skewed slightly to the right, with mean 4.7 cm and SD 2.1 cm.

a. If you select 50 books at random from this library, what is the probability that they will fit on a shelf that is 240 cm in length?

b. Does your answer to part a imply that the 50 books in the philosophy section probably will fit on one 240-cm shelf?

E51. The magazine Forbes found that the mean age of the chief executive officers (CEOs) of America’s 500 largest companies is about 55.7 years, with standard deviation 6.8 years. The ages are roughly mound-shaped, as shown in Display 7.51.

Display 7.51 Frequency histogram for ages of CEOs.

a. What do you find interesting about this histogram?

b. If you select ten of these CEOs at random, what is the probability that their average age is between 55 and 60?

E52. W. J. Youden (1900–71) weighed many new pennies and found that the distribution is approximately normal with mean 3.11 g and standard deviation 0.43 g. What are the reasonably likely weights of a roll of 50 randomly selected new pennies? [Source: W. J. Youden, Experimentation and Measurement (National Bureau of Standards, 1984), pp. 107–9.]

E53. Young men have an average height of about 68 in. with standard deviation 2.7 in. Young women have an average height of about 64 in. with standard deviation 2.5 in. Both distributions are approximately normal. A young man and a young woman are selected at random.

a. Describe the sampling distribution of the difference of their heights.

b. Find the probability that he is 2 in. or more taller than she is.

c. What is the probability that she is taller than he is?

E54. Two buses always arrive at a bus stop between 11:13 and 11:23 each day. Over a sample of 5 days, the average time between the times the two buses arrived was 7 minutes. Are the buses arriving at random in the 10-minute interval? Follow these steps to arrive at an answer you think is reasonable.

a. The population of interest is the set of possible times between the two arrivals (interarrival times). Describe a way to generate five of these interarrival times, assuming the two buses arrive at random and independent times during the 10-minute interval. Then take such a sample.

b. What was the mean interarrival time for your sample of size 5?

c. The dot plot in Display 7.52 shows the results of 100 repetitions of the process in parts a and b. What is the largest mean interarrival time recorded? Give an example of five pairs of arrival times that would give that mean.

Display 7.52 Dot plot of the mean difference between arrival times for two buses for samples of 5 days.

d. Based on the sampling distribution, do you think it’s reasonable to conclude that the buses in question were arriving at random?
E55. You select two digits at random from 0 through 9 (with replacement) and take their average. Your opponent also selects two digits at random and takes their average. You win an amount equal to the difference between your average and your opponent’s average. (If the difference is negative, you lose that amount.)

a. Describe how to simulate a distribution of the amount you win per play of the game. Do this simulation three times and record the differences.

b. Describe the distribution of all possible outcomes.

E56. In Chapter 4, you read about Kelly Acampora’s hamster experiment. The four hamsters raised in short days had enzyme levels

12.500 11.625 18.275 13.225

The hamsters raised in long days had enzyme levels

6.625 10.375 9.900 8.800

In this exercise, you’ll use a sampling distribution to decide whether this is persuasive evidence that hamsters raised in short days have higher enzyme levels than hamsters raised in long days. Your summary statistic for Kelly’s experiment will be $d$, the difference of the mean enzyme level of hamsters raised in short days and the mean enzyme level of hamsters raised in long days.

a. What is the value of the summary statistic $d$ for Kelly’s hamsters?

b. Suppose the length of a day makes no difference in enzyme levels, that is, suppose the eight numbers would have been the same if the hamsters had all received the opposite treatment. Use simulation to construct an approximate sampling distribution of all possible values of $d$. In other words, assign the hamsters at random so that four get each treatment but the enzyme level for the hamster is the same no matter what treatment it gets.

c. Is Kelly’s value of $d$ a reasonably likely outcome if the length of a day makes no difference? What can you conclude?

E57. Suppose you take a sample of size $n$ with replacement from a binomial population. As $n$ increases, describe what happens to the shape, mean, and standard error of the sampling distribution of

a. the sample proportion

b. the number of successes in the sample

E58. The histogram in Display 7.53 shows the number of airline passengers (in thousands) departing from or arriving at the 30 largest world airports during a recent year. The only outlier is Atlanta, with 83,606,583 passengers. The mean of this population is 42,520, and the standard deviation is 14,758.

The histograms and summary statistics in Display 7.54 show simulated distributions of 5000 sample means for samples of size 5, 10, and 20, selected without replacement from the numbers of passengers.

Check how well the shapes, means, and standard deviations of the simulated sampling distributions agree with what the theory says they should be. Do you see any reason why the theory you have learned should not work well in any of these cases?
Sample Size | Mean (in thousands) | Standard Deviation (in thousands) |
---|---|---|
5 | 42,296 | 6,085 |
10 | 42,441 | 3,790 |
20 | 42,526 | 1,909 |

E59. Bottle caps are manufactured so that their inside diameters have a distribution that is approximately normal with mean 36 mm and standard deviation 1 mm. The distribution of the outside diameters of the bottles is approximately normal, with mean 35 mm and standard deviation 1.2 mm. If a bottle cap and a bottle are selected at random (and independently), what is the probability that the cap is not too small to fit on the bottle?

E60. Refer to E59. Suppose a cap is too loose if it is at least 1.1 mm larger than the bottle. If a cap and a bottle are selected at random, what is the probability that the cap is too loose?

E61. A student drives to school in the morning and drives home in the afternoon. She finds that her commute time depends on the day of the week and whether it is morning or afternoon. Her data are shown in Display 7.55.

<table>
<thead>
<tr>
<th>Day</th>
<th>Morning Commute Time (in minutes)</th>
<th>Afternoon Commute Time (in minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monday</td>
<td>16</td>
<td>9</td>
</tr>
<tr>
<td>Tuesday</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>Wednesday</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>Thursday</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td>Friday</td>
<td>10</td>
<td>11</td>
</tr>
</tbody>
</table>

Display 7.55 Commute times.

a. What are the mean \( \mu \) and variance \( \sigma^2 \) of the morning commute times? Of the afternoon commute times?

b. If the student selects a day at random and finds the total commute time for that day, what are the mean and variance of the sampling distribution of this total commute time?

c. Are your answers in part b equal to the sum of those in part a? Explain why they should or shouldn’t be equal.
**AP Sample Test**

**AP1.** Five math teachers are asked how many pens they are currently carrying, and the results are 1, 1, 1, 2, 2. Random samples of size two are taken from this population (without replacement). What is the median of the sampling distribution of the median?

- A 1
- B 1.4
- C 1.5
- D 2
- E none of the above

**AP2.** With random sampling, which of the following is the best reason not to use the sample maximum as an estimator for the population maximum?

- A The sample maximum has too much variability.
- B The sample maximum is biased.
- C The sample maximum is difficult to compute.
- D The sample range is an unbiased estimator.
- E The sample maximum does not have a normally distributed sampling distribution.

**AP3.** In computing the sample standard deviation, the formula calls for a division by \( n - 1 \). Which of the following is the best reason for dividing by \( n - 1 \) instead of \( n \)?

- A For averages, always divide by \( n \), and for standard deviations, always divide by \( n - 1 \).
- B You use only \( n - 1 \) data values when computing standard deviations.
- C Dividing by \( n - 1 \) gives less variation in the sampling distribution of the population standard deviation.
- D Dividing by \( n - 1 \) makes the sample variance an unbiased estimator of the population variance in random sampling.
- E The sampling distribution of the sample standard deviation is closer to normal when dividing by \( n - 1 \).

**AP4.** The distribution of the population of the millions of household incomes in California is skewed to the right. Which of the following best describes what happens to the sampling distribution of the sample mean when the size of a random sample increases from 10 to 100?

- A Its mean gets closer to the population mean, its standard deviation gets closer to the population standard deviation, and its shape gets closer to normal.
- B Its mean stays constant, its standard deviation gets closer to the population standard deviation, and its shape gets closer to normal.
- C Its mean stays constant, its standard deviation gets smaller, and its shape gets closer to normal.
- D None of the above

**AP5.** The scores on a standardized test are normally distributed with mean 500 and standard deviation 110. In a randomly selected group of 100 test-takers, what is the probability that the mean test score is above 510?

- A less than 0.0001
- B 0.1817
- C 0.4638
- D 0.5362
- E 0.8183

**AP6.** A statistics teacher claims to be able to guess, with better than 25% accuracy, which of four symbols (circle, wavy lines, square, or star) is printed on a card. To test this claim, she guesses the symbol on 40 cards. If you use the normal approximation to the binomial to compute the probability that
she would get 15 or more (37.5%) correct just by random guessing, which of the following is closest to your z-score?

A 0.183
B 0.38
C 1.25
D 1.83
E 1067

AP7. In a large city, 20% of adults favor a new, stricter dress code for the local schools. Which of the following is closest to the probability that more than 24% of 1200 adults randomly selected for a survey will be in favor of the new dress code?

A less than 0.0001
B 0.0003
C 0.0029
D 0.38
E 0.9997

AP8. A study found that infants get a mean of 290 minutes of sleep between midnight and 6 a.m. with a standard deviation of 30 minutes. A parent wants to know the probability that two randomly selected infants will sleep a total of more than 600 minutes between midnight and 6 a.m., but a statistics student tells the parent that there's not enough information here to perform that calculation. Which of the following is the best reason for the student’s conclusion?

A The mean of the total cannot be determined.
B Statistical techniques can be used for the mean but not for the total.
C The standard deviation of the total cannot be determined.
D The shape of the sampling distribution of the total is unknown.
E The value of z cannot be determined.

Investigative Tasks

AP9. On an average day in 2004, about 246,000 vehicles traveled east on the Santa Monica freeway in Los Angeles to the interchange with the San Diego freeway. Assume that a randomly selected vehicle is equally likely to go straight through the interchange, go south on the San Diego freeway, or go north on the San Diego freeway. [Source: Caltrans, www.dot.ca.gov.]

a. What is the best estimate of the number of vehicles that will go straight through the interchange on an average day?

b. What numbers of vehicles are reasonably likely to go straight through?

c. On an average day, Caltrans found that 138,300 vehicles went straight through the interchange, 46,800 went north on the San Diego Freeway, and 60,900 went south. What can you conclude?

AP10. You estimate that the people using an elevator in an office building have an average weight of roughly 150 lb with a standard deviation of approximately 20 lb. The elevator is designed for a 2000-lb weight maximum. This maximum can be exceeded on occasion but should not be exceeded on a regular basis. Your job is to post a sign in the elevator stating the maximum number of people to ensure safe use. Keep in mind that it is inefficient to make this number too small but dangerous to make it too large.

a. What number would you use for maximum occupancy? Explain your reasoning and assumptions.

b. Otis Elevator Company sells the Holed Hydraulic Elevator, which lists a capacity of 2000 lb, or 13 people in the United States and 12 people in Canada. Are these numbers close to the number you decided on in part a? Explain why you would or would not expect that to be the case. [Source: www.otis.com, 2006.]
63% of U.S. households answering a recent survey reported that they own a pet, up from 56% in 1994, the year the survey started tracking this information. Are you confident that there has been an increase in the percentage of U.S. households that own a pet? Your answer will depend on how many households were included in the surveys.
Open the daily newspaper, a newsmagazine, or your own favorite magazine and you are likely to see the results of a public opinion poll. For example, a recent national survey found that 55% of singles ages 18–29 say that they aren’t in a committed relationship and are not actively looking for a romantic partner. This survey had a margin of error of 3%. [Source: Pew Research Center, February 13, 2006, Not Looking for Love: Romance in America, www.pewresearch.org.]

In this chapter, you will learn the two basic techniques of statistical inference: confidence intervals and significance testing. If you are familiar with polls, you have seen the idea of confidence interval. The survey’s result, 55% ± 3%, is equivalent to a confidence interval of 52% to 58%, an interval of plausible values for the true percentage. In your work in Chapter 1 on the Westvaco case, you used the idea behind significance testing when you saw that if you randomly selected three of the ten hourly workers to be laid off, you would be unlikely to get a group with an average age as large as the average age of those actually laid off by Westvaco.

The ideas on confidence intervals and significance testing developed here are fundamental to work in statistical inference. They will form the basis of the inferential procedures in the chapters to follow.

**In this chapter, you will learn**

- to construct and interpret a confidence interval to estimate the proportion of successes in a binomial population
- to use a significance test (hypothesis test) to decide if it is reasonable to conclude that your sample might have been drawn from a binomial population with a specified proportion of successes
- to construct and interpret a confidence interval for the difference between the proportion of successes in one population and the proportion of successes in another population
- to use a test of significance to decide if it is reasonable to conclude that two samples might have been drawn from two binomial populations that have the same proportion of successes
- to construct and interpret confidence intervals and tests of significance for experiments
Estimating a Proportion with Confidence

The Pew Research Center survey reported on the previous page found that 55% of singles ages 18–29 say that they aren’t in a committed relationship and are not actively looking for a romantic partner. This percentage is based on interviews with 1068 singles. The survey reported a margin of error of 3%.

The Pew Research Center didn’t ask all young singles in the United States, only 1068. However, unless there were some special difficulties in taking a random sample or with the wording of the question, the researchers are 95% confident that the error in the percentage is less than 3% either way. That is, they are 95% confident that if they were to ask all young singles in the United States, 55% ± 3%, or between 52% and 58%, would report that they aren’t in a committed relationship and are not actively looking for a romantic partner. What, exactly, can they mean by this? The answer involves the concept of reasonably likely events.

Reasonably Likely Events

In Section 7.3, you learned how much variability to expect in the proportions of successes \( \hat{p} \) from repeated random samples taken from a given binomial population. About 95% of all sample proportions \( \hat{p} \) will fall within about two standard errors of the population proportion \( p \), that is, within the interval

\[
p \pm 1.96 \sqrt{\frac{p(1-p)}{n}}
\]

where \( n \) is the sample size. The sample proportions in this interval are called reasonably likely. (Recall that this rule works well only under the condition that both \( np \) and \( n(1-p) \) are at least 10.)

Reasonably Likely Events and Rare Events

Reasonably likely events are those in the middle 95% of the distribution of all possible outcomes. The outcomes in the upper 2.5% and lower 2.5% of the distribution are rare events—they happen, but rarely.

Example: Reasonably Likely Results from Coin Flips

Suppose you will flip a fair coin 100 times. What are the reasonably likely values of the sample proportion \( \hat{p} \)? What numbers of heads are reasonably likely?
Here the probability of a success \( p \) on each trial is 0.5, and you have 100 trials—a sample size of 100. From the previous formula, 95% of all sample proportions \( \hat{p} \) should fall in the interval

\[
p \pm 1.96 \sqrt{\frac{p(1-p)}{n}} = 0.5 \pm 1.96 \sqrt{\frac{0.5(1-0.5)}{100}}
\]

or about 0.5 ± 0.1. So the reasonably likely values of \( \hat{p} \) are in the interval from 0.4 to 0.6. Values of \( \hat{p} \) outside this interval can happen, but they happen rarely.

Equivalently, in about 95% of the samples, the number of successes \( x \) in the sample will be in the interval

\[
xp \pm 1.96\sqrt{xp(1-p)} = 100(0.5) \pm 1.96\sqrt{100(0.5)(1-0.5)}
\]

or about 50 ± 10 successes.

Even though the Pew Research Center doesn’t know the value of \( p \), the actual proportion of young singles who aren’t in a committed relationship and are not actively looking for a romantic partner, it can use the idea in the preceding example. For each possible value of \( p \), Pew can compute how close to \( p \) most sample proportions will be. By knowing the variability expected in random samples taken from populations with different values of \( p \), Pew can estimate how close \( \hat{p} \) should be to the “truth.”

### Reasonably Likely Events

**D1.** Suppose 35% of a population think they pay too much for car insurance. A polling organization takes a random sample of 500 people from this population and computes the sample proportion, \( \hat{p} \), of people who think they pay too much for car insurance.

a. There is a 95% chance that \( \hat{p} \) will be between what two values?

b. Is the organization reasonably likely to get 145 people in the sample who think they pay too much for car insurance?

**D2.** In parts a–d, you will review the method of using simulation to construct a sampling distribution.

a. Using a calculator, computer, or random-digit table, take a random sample of size 40 from a population with 60% successes. Count the number of successes in your sample. [See Calculator Note 8A.]
b. Continue taking samples of size 40 until your class has the number of successes from 100 samples.

c. Plot your 100 values on a dot plot or a histogram.

d. From your plot in part c, what numbers of successes are reasonably likely?

e. Is your answer in part d close to what the theory predicts?

In Activity 8.1a, you will collect data about the proportion of students who can make the Vulcan salute. Later in this section, you will learn to find a confidence interval for the percentage of all students who can make the Vulcan salute.

**ACTIVITY 8.1a**

**The Vulcan Salute**

The Vulcan salute from Star Trek, which means “live long and prosper,” originated as part of a priestly blessing in Jewish ceremony. Your goal is to estimate what proportion of students can make this salute, clearly and easily, with both hands.

1. Without giving anyone the chance to practice, ask exactly 40 students to make the Vulcan salute with both hands at once. Count the number who are successful.

2. Suppose you can reasonably consider your sample of 40 students a random sample of all students. From the data you gathered, is it plausible that 10% of all students could make this salute? Is it plausible that the true percentage is 60%? What percentages do you think are plausible?

**The Meaning of a Confidence Interval**

In Activity 8.1b, you will make a chart that allows you to see the reasonably likely events for all population proportions, \( p \), when the sample size is 40. From there, it is a short step to confidence intervals.

**ACTIVITY 8.1b**

**Constructing a Chart of Reasonably Likely Events**

*What you’ll need: a copy of Display 8.1*

1. Suppose you take repeated random samples of size 40 from a population with 60% successes. What proportions of successes would be reasonably likely in your sample?

2. Compare the proportion of students you found able to make the Vulcan salute in Activity 8.1a to your answer in step 1. Is it plausible to assume that 60% of all students can make the salute? Explain your reasoning.

(continued)
3. On a copy of the chart in Display 8.1, draw a horizontal line segment aligned with 0.6 on the vertical axis, labeled “Proportion of Successes in the Population.” The line segment should stretch in length over your interval from step 1.

Your instructor will give you one of the population proportions whose line segments are missing in Display 8.1.

4. Compute the reasonably likely sample proportions for your population proportion, \( p \).

5. On your copy of the chart in Display 8.1, draw a horizontal line segment showing the reasonably likely sample proportions for your group’s population proportion, \( p \).

6. Get the reasonably likely sample proportions from the other groups in your class and complete the chart with line segments for those values of \( p \).

7. Refer to your results from Activity 8.1a. For which population proportions is your sample proportion reasonably likely?

The line segments you drew on your copy of Display 8.1 in Activity 8.1b show reasonably likely sample proportions when taking random samples of size 40 from a population with a given proportion of successes \( p \).

**Example: Plausible Percentages of Vulcan Saluters**

In a group of 40 adults doing Activity 8.1a, exactly 30 were able to make the Vulcan salute. Assuming this can be considered a random sample of all adults, is it plausible that if you tested all adults, you would find that 50% can make the salute? Is it plausible that 80% can make the salute? What percentages are plausible?
Solution

No, it’s not plausible that 50% of all adults can make the Vulcan salute, because 30 out of 40 (or \( \hat{p} = 0.75 \)) isn’t a reasonably likely sample proportion for a random sample taken from a population with 50% successes. You can see this from your chart or from Display 8.2 by going to the horizontal line segment next to the population proportion 0.5. This line segment goes from about 0.35 to 0.65, and that interval does not include a sample proportion, \( \hat{p} \), of 0.75. Of course, it is possible that the true percentage is 50% and that the adults were an unusual sample, but it’s not very plausible.

Is it plausible that 80% is the actual percentage of all adults who can make the salute? The horizontal line segment for a population percentage of 80% goes from 0.65 to 0.95, and that interval does include the sample proportion, \( \hat{p} \), of 0.75. If the true percentage is 80%, you are reasonably likely to get 30 adults who can salute Vulcan-style from a random sample of 40 adults.

To find the plausible percentages from your chart, draw a vertical line upward from \( \hat{p} = 0.75 \), as in Display 8.2. The line segments it crosses represent the plausible population percentages: 60%, 65%, 70%, 75%, 80%, and 85%. In all these populations, a sample proportion, \( \hat{p} \), of 0.75 would be a reasonably likely result.

The plausible percentages for the population proportion listed in the preceding example (60% to 85%) are called the 95% confidence interval for the percentage of all adults who can make the Vulcan salute. That is, you are 95% confident that if you were to test all adults, the percentage who can make the Vulcan salute would be somewhere in the interval from about 60% to 85%.
95% Confidence Interval for a Population Proportion \( p \)

A 95% confidence interval consists of those population proportions \( p \) for which the sample proportion \( \hat{p} \) is reasonably likely.

**The Meaning of a Confidence Interval**

Use your completed chart or the one in Display 8.2 to answer these questions.

D3. According to the 2000 U.S. Census, about 60% of Hispanics in the United States are of Mexican origin. Would it be reasonably likely that in a survey of 40 randomly chosen Hispanics, 27 are of Mexican origin? [Source: www.census.gov.]

D4. According to the 2000 U.S. Census, about 30% of people over age 85 are men. In a random sample of 40 people over age 85, would you be reasonably likely to get 60% who are men? [Source: www.census.gov.]

D5. Suppose that in a random sample of 40 toddlers, 34 know what color Elmo is. What is the 95% confidence interval for the percentage of toddlers who know what color Elmo is?

D6. Polls usually report a margin of error. Suppose a poll of 40 randomly selected statistics majors finds that 20 are female. The poll reports that 50% of statistics majors are female, with a margin of error of 15%. Use your completed chart to explain where the figure 15% came from.

D7. Why do we say we want a confidence interval for \( p \) rather than saying we want a confidence interval for \( \hat{p} \)?

**From the Chart to a Formula**

If you have noticed that Display 8.2 can be used to find a confidence interval only when the sample size is 40, you might be thinking, “Constructing a new chart each time they give me a different sample size is going to be a lot of work—perhaps it’s not too late to transfer to that pottery class.” But don’t leave yet. There is a quick and easy formula you can use.

To develop this formula, think again about the geometry behind a confidence interval. Suppose that in a random sample of size 40 the sample proportion, \( \hat{p} \), turns out to be 0.6. As shown in Display 8.3, you can approximate the 95% confidence interval by drawing a vertical line from \( \hat{p} = 0.6 \) and seeing which horizontal line segments it intersects. These horizontal line segments give the population proportions, \( p \), for which the sample proportion \( \hat{p} \) is a reasonably likely result. These populations are \( p = 0.45 \) up through \( p = 0.70 \) almost to \( p = 0.75 \). This is the confidence interval, the set of plausible values for \( p \).
Display 8.3 Finding a 95% confidence interval for \( p \) when the sample proportion, \( \hat{p} \), is 0.6 and the sample size, \( n \), is 40.

How can you find this confidence interval without the chart? The key is to notice that the two bold line segments in Display 8.4 intersect at the point (0.6, 0.6) and are approximately the same length. (They are approximately the same length because the two sets of endpoints of the horizontal line segments almost lie on two parallel lines with slope 1.)

Display 8.4 The length of the 95% confidence interval (the vertical line segment) is the same as the length of the horizontal line segment of reasonably likely sample proportions for \( p = 0.6 \).
The bold horizontal line segment represents the reasonably likely sample proportions, or the middle 95% of all sample proportions, from a population with \( p = 0.6 \). You know how to find its endpoints:

\[
p \pm 1.96 \sqrt{\frac{p(1-p)}{n}} = 0.6 \pm 1.96 \sqrt{\frac{0.6(1-0.6)}{40}} \approx 0.6 \pm 0.15
\]

The bold vertical line segment is the confidence interval, and it has the same endpoints—namely, 0.6 ± 0.15, or 0.45 and 0.75. So, to get the (vertical) confidence interval, all you have to do is find the endpoints of the horizontal line segment by substituting the known value of \( \hat{p} \) for the unknown value of \( p \) in the formula.

**From the Chart to a Formula**

D8. Suppose you observe a sample proportion, \( \hat{p} \), of 0.5 in a random sample of size 40.

  a. Use Display 8.2 to find an approximate 95% confidence interval for the population proportion.
  b. On a copy of Display 8.2, sketch in the two bold line segments for the proportion 0.5, as in Display 8.4.
  c. Explain the meaning of the bold horizontal line segment. Use a formula to find its endpoints.
  d. Explain the meaning of the bold vertical line segment. Find the endpoints of the bold vertical line segment using your result from part c.
  e. What are the endpoints of the 95% confidence interval?

D9. Explain the difference between the terms “reasonably likely sample proportions” and “plausible population proportions” in explaining the relationship between sample proportions and population proportions.

**Using the Formula**

You can use the reasoning described above to find a confidence interval for the percentage of successes \( p \) in a population, no matter what the sample size or the level of confidence desired. The general formula and the conditions for using it are given in this box.

**A Confidence Interval for a Population Proportion**

A confidence interval for the proportion of successes \( p \) in the population is given by the formula

\[
\hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
\]

Here \( n \) is the sample size and \( \hat{p} \) is the proportion of successes in the sample. The value of \( z^* \) depends on how confident you want to be that \( p \) will be in the

\( (continued) \)
(continued) confidence interval. If you want a 95% confidence interval, use $z^* = 1.96$; for a 90% confidence interval, use $z^* = 1.645$; for a 99% confidence interval, use $z^* = 2.576$. Table A on page 824 or your calculator will give you the value of $z^*$ for other confidence levels. [See Calculator Note 8B.]

This confidence interval is reasonably accurate when three conditions are met:

- The sample was a simple random sample from a binomial population.
- Both $np$ and $n(1 - \hat{p})$ are at least 10.
- The size of the population is at least 10 times the size of the sample.

The first two conditions listed in the box are necessary for you to be able to use the normal distribution (and $z$-scores) as an approximation to the binomial distribution. If the third condition isn’t met, your confidence interval will be longer than it needs to be.

### Margin of Error

The quantity

$$E = z^* \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

is called the **margin of error**. It is half the width of the confidence interval.

#### Example: Safety Violations

Suppose you have a random sample of 40 buses from a large city and find that 24 buses have a safety violation. Find the 90% confidence interval for the proportion of all buses that have a safety violation.

**Solution**

You will check the three conditions in D10.

For a 90% confidence interval, use $z^* = 1.645$. The confidence interval for the proportion $p$ of all buses that have a safety violation based on a sample proportion, $\hat{p}$, of $\frac{24}{40}$ or 0.6 can be written

$$\hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = 0.6 \pm 1.645 \sqrt{\frac{0.6(1 - 0.6)}{40}}$$

$$\approx 0.6 \pm 1.645(0.077)$$

$$\approx 0.6 \pm 0.13$$

or about (0.47, 0.73).
You are 90% confident that the percentage of all of this city’s buses that have a safety violation is between 47% and 73%. The margin of error for this survey is 0.13, or 13%.

You can also use a calculator to calculate confidence intervals. [See Calculator Note 8C.] Shown here are the values for the previous example.

<table>
<thead>
<tr>
<th>I-PropZInt</th>
<th>I-PropZInt</th>
</tr>
</thead>
<tbody>
<tr>
<td>x=24</td>
<td>(.47259, .72741)</td>
</tr>
<tr>
<td>n=40</td>
<td>p=6</td>
</tr>
<tr>
<td>C-Level:.90</td>
<td>n=40</td>
</tr>
</tbody>
</table>

**Using the Formula**

D10. In the bus example, are the three conditions for constructing a confidence interval met? Explain. (You should always verify that the three conditions are met before reporting a confidence interval.)

D11. What, in words, does $np$ represent in the bus example? What, in words, does $n(1 - \hat{p})$ represent in the bus example? What do these two quantities represent in general?

**The Capture Rate**

Sometimes a confidence interval “captures” the true population proportion and sometimes it doesn’t. The capture rate of a method of constructing confidence intervals is the proportion of confidence intervals that contain the population parameter in repeated usage of the method. With a good method, the capture rate is equal to the level of confidence advertised. For example, if a polling company uses 95% confidence intervals in a large number of different surveys, the population proportion $p$ should be in 95% of them. You will test whether that is indeed the case in Activity 8.1c (and explore this further in Investigative Tasks AP9 and AP10 on pages 558–559).

**ACTIVITY 8.1c**

**The Capture Rate**

**What you’ll need:** a calculator or one table of random digits per student, a copy of Display 8.5

1. Your instructor will assign you a group of 40 random digits, or you will generate 40 random digits on your calculator.
2. Count the number of even digits in your sample of 40 random digits.
3. Use the formula to construct a 95% confidence interval for the proportion of random digits that are even.

(continued)
4. Each member of your class should draw his or her confidence interval as a vertical line segment on a copy of Display 8.5.

5. What is the true proportion of all random digits that are even?


Student: In the activity, about 95% of the 95% confidence intervals captured the true population proportion of 0.5. That was no surprise. But I don't see why that happened. Just calling something a 95% confidence interval doesn't make it one.

Statistician: You're right. This isn't obvious. For me to explain the logic to you, you'll have to answer some questions as we go along.

Student: Okay.

Statistician: Go back and ask those of your classmates who had confidence intervals that captured the true proportion of even digits \( p = 0.5 \) what values of \( \hat{p} \) they got. We'll talk again tomorrow.

Student: (The next day) They had values of \( \hat{p} \) between 0.35 and 0.65.

Statistician: Right again. Now look at Display 8.6. What do you notice about these values of \( \hat{p} \)?

Student: They all lie on the horizontal line segment for \( p = 0.5 \).

Statistician: Yes, the values of \( \hat{p} \) that give a confidence interval that captures \( p = 0.5 \) are the reasonably likely outcomes for \( p = 0.5 \). What is the chance of getting one of these “good” values of \( \hat{p} \)?

Student: 95%!
Display 8.6  The population proportion $p$ is in the confidence interval if and only if $\hat{p}$ is a reasonably likely outcome for $p$.

It is correct to say that you expect the true value of $p$ to be in 95 out of every 100 of the 95% confidence intervals you construct. However, it is not correct to say, after you have found a confidence interval, that there is a 95% probability that $p$ is in that confidence interval. Here is an example that shows why. If you pick a date in the next millennium at random, it is reasonable to say that there is a $\frac{1}{7}$ chance you will pick a Tuesday. However, suppose the date you pick turns out to be March 3, 3875. It sounds a bit silly to say that there is a $\frac{1}{7}$ chance that March 3, 3875, is a Tuesday. Once the date is selected, there is no randomness left. Either March 3, 3875, is a Tuesday or it isn’t. All you can say is that the process you have used to select a date gives you a Tuesday $\frac{1}{7}$ of the time. (This example is credited to Wes White, an AP Statistics teacher in Los Angeles.)

**DISCUSSION**

**The Capture Rate**

D12. Was every student’s confidence interval in Activity 8.1c the same? Is this how it should be? Explain.

D13. After working through parts a and b, explain the relationship, if any, between the length of the confidence interval and the capture rate.

a. Construct an example of two confidence intervals with the same width but different capture rates (one higher than the other).

b. Construct examples of two confidence intervals with the same capture rate but different width.

**Margin of Error and Sample Size**

You might have noticed that the 95% confidence intervals for large sample sizes are narrower than those for small sample sizes. This makes sense—the larger the sample size, the closer $\hat{p}$ should be to $p$. Let’s see how this works for a specific example.
Example: The Effect of Sample Size on the Margin of Error

Suppose you take a random sample and get \( \hat{p} = 0.7 \).

a. If your sample size is 100, find the 95% confidence interval for \( p \), and state the margin of error.

b. What happens to the confidence interval and margin of error if you quadruple the sample size, to 400?

Solution

a. The 95% confidence interval for a sample size of 100 is

\[
\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = 0.7 \pm 1.96 \sqrt{\frac{0.7(1 - 0.7)}{100}}
\]

\[
\approx 0.7 \pm 1.96(0.0458)
\]

\[
\approx 0.7 \pm 0.0898
\]

The 95% confidence interval is (0.6102, 0.7898), and the margin of error is approximately 0.0898.

b. If you quadruple your sample size to 400, your 95% confidence interval would be

\[
\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = 0.7 \pm 1.96 \sqrt{\frac{0.7(1 - 0.7)}{400}}
\]

\[
\approx 0.7 \pm 1.96(0.0229)
\]

\[
\approx 0.7 \pm 0.0449
\]

The 95% confidence interval is (0.6551, 0.7449), and the margin of error is approximately 0.0449, or half what it was before.

DISCUSSION

Margin of Error and Sample Size

D14. All else being equal, will the margins of error for samples of size 80 be larger or smaller than those for samples of size 40? Explain your reasoning.

D15. You can use Display 8.2 only for samples of size 40 and for 95% confidence intervals.

a. How would a chart for 99% confidence intervals be different from Display 8.2?

b. How would the sizes of the margins of error be different?

What Sample Size Should You Use?

People who conduct surveys often ask statisticians, “What sample size should I use?” The simple answer is that the larger the sample size, \( n \), the more precise the results will be. When everything else is held constant, the margin of error is smaller with a larger sample size than with a smaller one. However, researchers...
always have limited time and money, so for practical reasons they have to limit the size of their samples. For a confidence interval, the margin of error, \( E \), is approximately

\[
E = z^* \cdot \sqrt{\frac{p(1 - p)}{n}}
\]

You can solve this formula for \( n \):

\[
E^2 = (z^*)^2 \cdot \frac{p(1 - p)}{n}
\]

\[
n = (z^*)^2 \cdot \frac{p(1 - p)}{E^2}
\]

To use this formula, you have to know the margin of error, \( E \), that is acceptable. You have to decide on a level of confidence so you know what value of \( z^* \) to use, and you have to have an estimate of \( p \). If you have a good estimate of \( p \), use it in this formula. If you do not have a good estimate of \( p \), use \( p = 0.5 \). Using this value for \( p \) might give you a sample size that is a bit too large, but it will never give one that is too small. (This is because the largest possible value of \( p(1 - p) \) occurs when \( p = 0.5 \). See E21.)

**Example: Estimating the Needed Sample Size**

What sample size should you use for a survey if you want the margin of error to be at most 3% with 95% confidence but you have no estimate of \( p \)?

**Solution**

Because you do not have an estimate of \( p \), use \( p = 0.5 \). Then

\[
n = (z^*)^2 \cdot \frac{p(1 - p)}{E^2} = 1.96^2 \cdot \frac{0.5(1 - 0.5)}{0.03^2}
\]

\[
= 1067.111
\]

Because you must have a sample size of at least 1067.111, round up to 1068. To get a margin of error around 3% is one reason why national polling organizations use sample sizes of around 1000.

**DISCUSSION**

**What Sample Size Should You Use?**

D16. Examine the formula for a confidence interval and explain why it is true that when everything else is held constant, the margin of error is smaller with a larger sample size than with a smaller one.

D17. Suppose it costs $5 to survey each person in your sample. You judge that \( p \) is about 0.5. What will your survey cost if you want a 95% confidence interval with a margin of error of about 10%? About 1%? About 0.1%?
Back to the Survey of Young Singles

The beginning of this section told about a recent Pew Research Center survey reporting that 55% of singles ages 18–29 aren’t in a committed relationship and aren’t actively looking for a romantic partner. The sample size was 1068. This survey had a margin of error of 3%, so the 95% confidence interval was 52% to 58%.

You now should be able to answer questions like these about this confidence interval:

- **What is it that you are 95% sure is in the confidence interval?**
  The proportion of all singles ages 18–29 in the United States who aren’t in a committed relationship and aren’t actively looking for a romantic partner.

- **What is the interpretation of the confidence interval of 52% to 58%?**
  You are 95% confident that if you could ask all singles ages 18–29 if they aren’t in a committed relationship and aren’t actively looking for a romantic partner, between 52% and 58% would say yes.

- **What is the meaning of “95% confidence”?**
  If you were to take 100 random samples of young singles and compute the 95% confidence interval from each sample, then you can expect 95 of the confidence intervals to contain the proportion of all young singles ages 18–29 who aren’t in a committed relationship and aren’t actively looking for a romantic partner.

- **What populations plausibly could have resulted in a sample proportion of 0.55?**
  A 95% confidence interval consists of those population proportions for which the observed sample proportion is a reasonably likely event. In this case, if the population proportion were anything from 0.52 to 0.58, you would be reasonably likely to get a sample proportion of 0.55 in a sample of size 1068.

DISCUSSION

Back to the Survey of Young Singles

D18. Verify the margin of error for the Pew Research Center survey.

D19. Sometimes polls report the “error attributable to sampling” instead of the “margin of error.” Explain what this means.

D20. Most legitimate polling organizations publish details on how their polls are conducted. Many include a statement similar to this: “In addition to sampling error, the practical difficulties of conducting a survey of public opinion may introduce other sources of error.” List at least three other possible sources of error and explain why they are not included in the sampling error measured by statistical formulas.

Summary 8.1: Estimating a Proportion with Confidence

A 95% confidence interval consists of those population percentages $p$ for which the observed sample proportion $\hat{p}$ is reasonably likely.

The formula for a confidence interval has the form

$$\text{statistic} \pm \text{(critical value)} \cdot \text{(standard deviation of statistic)}$$
In the case of estimating a proportion, the confidence interval is given by

\[ \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \]

where \( \hat{p} \) is the observed sample proportion and \( n \) is the sample size.

With \( z^* = 1.645 \), the method will capture the population proportion 90% of the time; with \( z^* = 1.96 \), it will do so 95% of the time; and with \( z^* = 2.576 \), it will do so 99% of the time.

This confidence interval is reasonably accurate when three conditions are met:

- The sample was a simple random sample from a binomial population.
- Both \( n\hat{p} \) and \( n(1 - \hat{p}) \) are at least 10.
- The size of the population is at least 10 times the size of the sample.

There are two parts to giving an interpretation of a confidence interval: describing what is in the confidence interval and describing what is meant by "confidence." For a 95% confidence interval, for example, this interpretation would include:

- You are 95% confident that, if you could examine each unit in the population, the proportion of successes, \( p \), in this population would be in this confidence interval.
- If you were able to repeat this process 100 times and construct the 100 resulting confidence intervals, you’d expect 95 of them to contain the population proportion of successes, \( p \). In other words, 95% of the time the process results in a confidence interval that captures the true value of \( p \).

Remember that there can be sources of error other than sampling error. For example, if the samples aren’t randomly selected or if there is a problem such as a poorly worded questionnaire, the capture rates don’t apply.

To estimate the sample size, \( n \), needed for a given margin of error \( E \), use the formula

\[ n = (z^*)^2 \cdot \left( \frac{\hat{p}(1 - \hat{p})}{E^2} \right) \]

with a rough estimate of \( \hat{p} \) if you have it; if not, use \( \hat{p} = 0.5 \).

**Practice**

**Reasonably Likely Events**

P1. Suppose 40% of students in your graduating class plan to go on to higher education. You survey a random sample of 50 of your classmates and compute the sample proportion \( \hat{p} \) of students who plan to go on to higher education.

a. There is a 95% chance that \( \hat{p} \) will be between what two numbers?

b. Is it reasonably likely to find that 25 students in your sample plan to go on to higher education?

P2. Describe how to use simulation to find the reasonably likely sample proportions for a random sample of size 40 taken from a population with \( p = 0.3 \).

P3. According to the U.S. Census Bureau, about 16% of the residents of the country do not have health insurance. Suppose a polling agency randomly selects 200 residents. What numbers of residents without health insurance are reasonably likely?
The Meaning of a Confidence Interval

Use Display 8.2 on page 472 to answer P4–P8.

P4. Suppose you flip a fair coin 40 times. How many heads is it reasonably likely for you to get?

P5. About 65% of 18- and 19-year-olds are enrolled in school. If you take a random sample of 40 randomly chosen 18- and 19-year-olds, would you be reasonably likely to find that 33 were in school? [Source: U.S. Census Bureau, Statistical Abstract of the United States, 2006, Table 209.]

P6. About 85% of people in the United States age 25 or over have graduated from high school. In a random sample of 40 people age 25 or older, how many high school graduates are you reasonably likely to get? What proportions of high school graduates are you reasonably likely to get? [Source: U.S. Census Bureau, Statistical Abstract of the United States, 2006, Table 214.]

P7. In a random sample of 40 adults, 25% know what color Elmo is. What is the 95% confidence interval for the percentage of all adults who know what color Elmo is?

P8. Suppose that in a random sample of 40 retired women, 45% of the women travel more than they did while they were working. Find the 95% confidence interval for the proportion of all retired women who travel more.

From the Chart to a Formula

P9. Suppose that in a random sample of 40 students from your school, 25 are wearing sneakers. Find the 95% confidence interval for the percentage of all students who wear sneakers, a. using the chart in Display 8.2 b. without using the chart

Using the Formula

P10. In a survey consisting of a randomly selected national sample of 600 teens ages 13–17, conducted from July 6 to September 4, 2005, 65% responded that their school was helping them discover what type of work they would love to do as a career. [Source: Gallup, Teens: Schools Help Students Find Career Path, poll.gallup.com.] a. Check to see if the three conditions for computing a confidence interval are met in this case. b. Find a 95% confidence interval for the percentage of all teens in the United States who would respond that their school is helping them discover a career path. What is the margin of error? c. Find a 90% confidence interval for the percentage of all teens in the United States who would respond that their school is helping them discover a career path. What is the margin of error? d. Which confidence interval, 90% or 95%, is wider? Why should that be the case?

P11. In the same survey as in P10, 4% of the 600 students responding gave their school a D rating (on a scale of A, B, C, D, F). a. Check to see if the three conditions for computing a confidence interval are met in this case. b. Find a 95% confidence interval for the percentage of all teenagers in the United States who would give their school a D rating. c. How does the width of the confidence interval in part b compare to that of the confidence interval in part b of P10? What is the reason for this?

The Capture Rate

P12. Suppose you know that a population proportion, \( p \), is 0.60. Now suppose 80 different students are going to select independent random samples of size 40 from this population. Each student constructs his or her own 90% confidence interval. How many of the resulting confidence intervals would you expect to include the population proportion, \( p \), of 0.60?
P13. A quality-control plan in a plant that manufactures computer mice calls for taking 20 different random samples of mice during a week and estimating the proportion of defective mice in a 95% confidence interval for each sample. If the percentage of defectives is maintained at 4% throughout the week, how many of the confidence intervals would you expect not to capture this value?

Margin of Error and Sample Size

P14. All else being equal, how would a chart for samples of size 100 be different from Display 8.2? How would the widths of the confidence intervals be different?

P15. What value of $z^*$ should you use for an 80% confidence interval? Will the margin of error for 80% confidence be longer or shorter than for 95% confidence? Explain your answer.

P16. Suppose you take a survey with sample size $n$ and get a margin of error that is three times larger than you would like. What sample size should you have used to obtain the desired margin of error? Justify your answer.

What Sample Size Should You Use?

P17. The accuracy of opinion polls is examined closely in presidential election years. Suppose your large polling organization is planning a final poll before the presidential election.

a. What sample size should you use to have a margin of error of 2% with a 95% confidence level?

b. What sample size should you use to have a margin of error of 1% with a 99% confidence level?

c. What sample size should you use if you want a margin of error of 0.5% with 90% confidence?

P18. A Gallup poll found that 29% of adult Americans report that drinking has been a source of trouble in their families. Gallup asks this question every year. What sample size should Gallup use next year to get a margin of error of 3% with 95% confidence and survey as few people as possible? [Source: Gallup News Service, July 31, 2006.]

Back to the Survey of Young Singles

P19. A recent survey about the use of the Internet among teens gave this information: 76% of middle school and high school students who have Internet access go online to get news or information about current events. Of students ages 12–17, the study found that 87%, or about 20 million, have Internet access. From that population, the study obtained 971 responses to the question about using the Internet to get news. [Source: Pew Internet, 2005, www.pewinternet.org.]

a. Check that the three conditions are met for computing a confidence interval for the percentage of students with Internet access who go online to get news or information about current events.

b. Regardless of your answer to part a, compute the 95% confidence interval.

c. What is it that you are 95% sure is in the confidence interval?

d. Interpret this confidence interval.

e. Explain the meaning of “95% confidence.”
E1. Review the use of the chart in Display 8.2 on page 472 by answering these questions.
   a. The proportion of successes in a random sample of size 40 is 0.90. Is this sample proportion reasonably likely if the population has 75% successes?
   b. In the multicandidate 2000 presidential election, Al Gore received 48% of the votes cast, “winning” the popular vote but losing in the electoral college. In a random sample of 40 voters in this election, what is the largest proportion of people who voted for Vice President Gore that is reasonably likely? The smallest proportion?
   c. Suppose a poll of a random sample of 40 shoppers finds that 16 prefer open-air malls to enclosed malls. What is the 95% confidence interval for the percentage of all shoppers who prefer open-air malls?

E2. Review the use of the chart in Display 8.2 by answering these questions.
   a. The proportion of successes in a random sample is 0.30. Is this sample proportion reasonably likely in a sample of size 40 if the population proportion is 0.40?
   b. According to the U.S. Census Bureau, about 16% of the residents of the country do not have health insurance. In a random sample of 40 residents, what is the largest proportion of uninsured people that is reasonably likely? The smallest proportion?
   c. Suppose a poll of a random sample of 40 shoppers finds that 16 prefer open-air malls to enclosed malls. What is the 95% confidence interval for the percentage of all shoppers who prefer open-air malls?

E3. A bilge alarm warns a boat captain when water begins rising in the bottommost part of the boat (the “bilge”). A working bilge alarm is considered crucial in preventing the flooding of fishing boats. Only 15% of 40 boats sampled from the ocean-fishing fleet in the United Kingdom had bilge alarms that were placed conveniently for testing and maintenance. [Source: Banff & Buchan College and the Universities of Glasgow and Strathclyde, Flooding of UK Fishing Boats, January 2003.]
   a. Find an approximate 95% confidence interval, using Display 8.2, for the proportion of all ocean-fishing boats in the United Kingdom with appropriately placed bilge alarms.
   b. Find a 95% confidence interval using the formula on page 475.
   c. Compare the answers in parts a and b. Give reasons for any discrepancy.

E4. Suppose a random sample of size 40 produces a sample proportion, \( \hat{p} \), of 0.80.
   a. Find an approximate 95% confidence interval from Display 8.2.
   b. Find a 95% confidence interval using the formula on page 475.
   c. Compare the answers in parts a and b. Give reasons for any discrepancy.

E5. A nationwide survey of 19,441 teens about their attitudes and behaviors toward epilepsy found that only 51% knew that epilepsy is not contagious.

The executive summary from the Epilepsy Foundation for this survey gave this technical information:

The two-page survey was distributed to teens nationally by 20 affiliates of the Epilepsy Foundation from March 2001 through July 2001 in schools selected by each affiliate. A total of 19,441 valid surveys were collected. Mathew Greenwald & Associates, Inc., edited the surveys and performed the data entry. Greenwald & Associates was also responsible for the tabulation, analysis, and reporting of the data. The data were weighted by age and region to reflect national percentages.

The margin of error for this study (at the 95% confidence level) is plus or minus approximately 1%.

The survey was funded by the Centers for Disease Control and Prevention. [Source: www.efa.org, April 16, 2002.]
a. Do you have any reservations about the design of this study?

b. Are the conditions for constructing a confidence interval met?

c. Does a computation of the margin of error using the formula in this section agree with that given in the press release?

d. Assuming the conditions to construct a confidence interval have been met, what is it that you are 95% sure is in the confidence interval?

e. Explain the meaning of “95% confidence.”

E6. Read this article.

Study: Teens Optimistic About Innovation
Brian Bergstein
Associated Press

BOSTON—Teenagers have some seemingly high expectations about what technology might bring over the next decade, according to a new Massachusetts Institute of Technology study.

For example, 33 percent of teens predicted that gasoline-powered cars will go the way of the horse and buggy by 2015. Just 16 percent of adults agreed. Meanwhile, 22 percent of teenagers predicted desktop computers will become obsolete a decade from now, while only 10 percent of adults agreed. Adults, on the other hand, were far more certain about the demise of the landline telephone by 2015 (45 percent made that prediction) than teenagers (17 percent) . . .

But he [the director of the survey] also wonders whether enough of today's teens are in a position to invent such solutions, noting that engineering was teens' third-most attractive career choice, picked by 14 percent as the field that most interested them—and just 4 percent of girls. Only 9 percent of all teens said they were leaning toward science. The top two career choices: Arts and medicine, each picked by 17 percent of the kids surveyed.

The Lemelson–MIT program, which focuses on encouraging young people to pursue innovation, commissioned its “invention index” in November, interviewing 500 teens and 1,030 adults nationwide. The margin of sampling error was plus or minus 4 percentage points for teens and 3 for adults. [Source: www.siliconvalley.com, June 11, 2006.]

a. What would you like to know about the survey design that is not described here?

b. Are the conditions for constructing confidence intervals met for teens’ responses? For adults’ responses?

c. Do the computations of the margins of error for both teens and adults, using the formula in this section, agree with that given in the press release?

d. Assuming the conditions have been met to construct a 95% confidence interval for the proportion of teenagers who think engineering is an attractive career choice, interpret this confidence interval.

e. Explain the meaning of “95% confidence.”

E7. A sample of 549 randomly selected teenagers ages 13–17 were asked whether it is appropriate for parents to install a computer program limiting what teens can access on the Internet. Fifty-two percent responded that, yes, this was an appropriate parental measure. Check the conditions and then construct a 90% confidence interval for the proportion of teenagers nationwide who would think this an appropriate measure.


E8. In a CBS News/New York Times poll, 885 adults nationwide were asked this question: “How worried are you that popular culture—that is television, movies, and music—is lowering the moral standards in this country: very worried, somewhat worried, not too worried, or not at all worried?”

Forty percent of those interviewed said that they were very worried. Check the conditions and then construct a 90% confidence interval for the proportion of adult Americans who are very worried that popular culture is lowering the moral standards of the country.

[Source: www.pollingreport.com.]
E9. Suppose a random sample of size 100 produces a sample proportion of 0.60. Can Display 8.2 be used in this case to find a 95% confidence interval estimate of the population proportion? Why or why not?

E10. A random sample of size 40 produces a sample proportion of 0.05. Can Display 8.2 be used in this case to produce a 95% confidence interval estimate of the population proportion? Why or why not?

E11. A U.S. News & World Report survey of 1000 adults selected from the general public (published April 15, 1996) reported that 81% thought TV contributes to a decline in family values. If the sample was randomly selected, what can you say about the proportion of all adults who think TV contributes to a decline in family values? (As always, discuss whether the three conditions are met, give the confidence interval itself, and give an interpretation of this interval, stating clearly what it is that you hope will be in the confidence interval.)

E12. Another part of the U.S. News & World Report survey went to Hollywood leaders because the magazine also wanted to see what Hollywood leaders thought. Of 6059 mailed surveys, only 570 were returned. Among the returned surveys, 46% thought TV contributed to a decline in family values. The magazine does not report a margin of error for this part of the survey. Should a margin of error be reported? Explain.

E13. If everything else in your sample is left unchanged,
   a. what happens to the width of a confidence interval as the sample size, n, increases?
   b. what happens to the width of a confidence interval as the degree of confidence you have in the answer increases?

E14. Explain the difference between \( p \) and \( \hat{p} \). Which is the parameter? Which is always in the confidence interval?

E15. The poll described in E7 reports that a random sample of 549 teenagers results in a margin of error of approximately 5%.
   a. Is this correct?
   b. What sample size should be used if the margin of error in estimating a proportion is to be only 3%?

E16. If you want a margin of error that is one-quarter of what you estimate it will be with your current sample size, by what factor should you increase the sample size?

E17. You want to determine the percentage of seniors who drive to school. You take a random sample of 125 seniors and ask them if they drive to school. Your 95% confidence interval turns out to be from 0.69 to 0.85. Select each correct interpretation of this situation. There might be no, one, or more than one correct statement.
   A. 77% of the seniors in your sample drive to school.
   B. 95% of all seniors drive to school from 69% to 85% of the time, and the rest drive more frequently or less frequently.
   C. If the sampling were repeated many times, you would expect 95% of the resulting samples to have a sample proportion that falls in the interval from 0.69 to 0.85.
   D. If the sampling were repeated many times, you would expect 95% of the resulting confidence intervals to contain the proportion of all seniors who drive to school.
   E. You are 95% confident that the proportion of seniors in the sample who drive to school is between 0.69 and 0.85.
   F. You are 95% confident that the proportion of all seniors who drive to school is in the interval from 0.69 to 0.85.
   G. All seniors drive to school an average of 77% of the time.
   H. A 90% confidence interval would be narrower than the interval given.
E18. A survey was conducted to determine what adults prefer in cell phone services. The results of the survey showed that 73% of the people wanted email service, with a margin of error of plus or minus 4%. Which of these sentences explains most accurately what is meant by “plus or minus 4%”?

A. They estimate that 4% of the population surveyed might change their minds between the time the poll is conducted and the time the survey is published.

B. There is a 4% chance that the true percentage of adults who want email service is not in the confidence interval from 69% to 77%.

C. Only 4% of the population was surveyed.

D. To get the observed sample proportion of 73% would be unlikely unless the actual percentage of all adults who want email service is between 69% and 77%.

E. The probability that the sample proportion is in the confidence interval is 0.04.

E19. In Section 5.2, you studied how to use tables of random digits to simulate probabilities. In the example on page 305, a simulation of 5000 runs estimated that the probability that both a randomly selected man and a randomly selected woman would wash their hands before leaving a public restroom was about 0.67.

a. What is the margin of error for this simulation?

b. How many runs of the simulation would you need to estimate this probability with a margin of error of 0.002?

de. Explain how to simulate the probability that the lines will be shut down.

d. Display 8.7 shows the number of defectives in each of 100 runs. (No values larger than 7 were observed.) Use these data to decide how many additional runs to make so that you can estimate the probability of seeing five or more defectives in the pooled output of the two lines with a margin of error of 0.01.

<table>
<thead>
<tr>
<th>Total Defects</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4</td>
<td>4.00</td>
</tr>
<tr>
<td>1</td>
<td>16</td>
<td>16.00</td>
</tr>
<tr>
<td>2</td>
<td>24</td>
<td>24.00</td>
</tr>
<tr>
<td>3</td>
<td>26</td>
<td>26.00</td>
</tr>
<tr>
<td>4</td>
<td>19</td>
<td>19.00</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>8.00</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>2.00</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Display 8.7 Results of 100 runs of a simulation.

E20. In a small factory, one assembly line produces about 10% defectives (products that need to be sent back for some reworking), and the neighboring line produces about 20% defectives. Ten items are randomly selected from each line, and the total number of defective items for the two lines together is counted. If this number is 5 or more, the lines are shut down for inspection. (It is impractical to shut down one line at a time in this small factory.)

a. Let $y = x(1-x)$, where $0 \leq x \leq 1$. Why let $y = x(1-x)$? Why restrict $x$ to $0 \leq x \leq 1$?

b. Sketch the graph of $y = x(1-x)$. What is the name of this type of function?

c. Where does the largest possible value of $y$ occur for this function? At what value of $x$ does this occur? What is the largest possible value of $y$?

E21. If you don’t have an estimate for $p$ when computing the needed sample size, you use $p = 0.5$. In this exercise, you will show that this is the safest value to use because $p = 0.5$ requires the largest sample size, all other things being equal. Answer parts a–c to prove that the largest possible value of $p(1-p)$ is achieved when $p = 0.5$.

a. Let $y = x(1-x)$, where $0 \leq x \leq 1$. Why let $y = x(1-x)$? Why restrict $x$ to $0 \leq x \leq 1$?

b. Sketch the graph of $y = x(1-x)$. What is the name of this type of function?

c. Where does the largest possible value of $y$ occur for this function? At what value of $x$ does this occur? What is the largest possible value of $y$?

E22. Explain why it is not appropriate to use the method of constructing a confidence interval for a population proportion described in this section when the condition that both $np$ and $n(1-\hat{p})$ are at least 10 is violated. You can use examples or a simulation to support your arguments.
E23. Supply a reason for each step in parts a–d in this alternative explanation of why the formula for a 95% confidence interval for a population proportion is reasonable.

a. Ninety-five percent of all sample proportions \( \hat{p} \) are within 1.96 standard errors of \( p \).

b. There is a 95% chance that the proportion \( \hat{p} \) from a specific random sample will be within 1.96 standard errors of \( p \).

c. There is a 95% chance that \( p \) will be within 1.96 standard errors of \( \hat{p} \).

d. Thus, the formula for a 95% confidence interval is approximately

\[
\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}
\]

E24. Observe 40 students on your campus and estimate, using 98% confidence, the percentage of students who carry backpacks.

### 8.2 Testing a Proportion

People often make decisions from data by comparing the results from a sample to some predetermined standard. These kinds of decisions are called tests of significance because the goal is to test the significance of the difference between the sample and the standard. If the difference is small, there is no reason to conclude that the standard doesn't hold. When the difference is large enough that it can't reasonably be attributed to chance, you can conclude that the standard no longer holds. In other words, the difference between the sample data and the standard is statistically significant.

Here is an example of this type of reasoning. About 2% of barn swallows have white feathers in places where the plumage is normally blue or red. The white feathers are caused by genetic mutations. In 1986, the Russian nuclear reactor at Chernobyl leaked radioactivity. Researchers continue to be concerned that the radiation might have caused mutations in the genes of humans and animals that have been passed on to offspring. In a sample of 266 barn swallows captured around Chernobyl between 1991 and 2000, about 16% had white feathers in places where the plumage is normally blue or red. Researchers compared the proportion, \( \hat{p} \), of 0.16 in the sample of captured barn swallows to the standard, 0.02. If the overall percentage was still only 2%, it is not reasonably likely that they would get 16% in the sample. So they came to the conclusion that there was an increased probability of genetic mutations in the Chernobyl area. [Source: A. P. Moller and T. A. Mousseau. “Albinism and Phenotype of Barn Swallows (Hirundo rustica) from Chernobyl,” Evolution 55 (2001): 2097–2104.]

In Activity 8.2a, you will record the results of 40 flips of a penny and 40 spins of a penny. You will use the data later to perform tests of significance.

### ACTIVITY 8.2a Spinning and Flipping Pennies

**What you’ll need:** pennies (at least one per student)

People tend to believe that pennies are balanced. They generally have no qualms about flipping a penny to make a fair decision. Is it really the case that penny flipping is fair? What about spinning pennies?

(continued)
1. Begin spinning the penny that your instructor provides. To spin, hold the penny upright on a table or the floor with the forefinger of one hand and flick the side edge with a finger of the other hand. The penny should spin freely, without bumping into things before it falls. Spin pennies until your class has a total of 40 spins. Count the number of heads and compute \( \hat{p} \).

2. Do you believe heads and tails are equally likely when spinning pennies, or can you reject this standard? Explain.

3. Begin flipping your penny. Let the penny fall, and record whether it lands heads or tails. Flip pennies until your class has a total of 40 flips. Count the number of heads and compute \( \hat{p} \).

4. Do you believe heads and tails are equally likely when flipping pennies, or can you reject this standard? Explain.

**Informal Significance Testing**

The logic involved in deciding whether to reject the standard that spinning a penny results in heads 50% of the time is the same as that used to construct a confidence interval.

**Example: Jenny and Maya’s Spins**

Jenny and Maya wonder if heads and tails are equally likely when a penny is spun. They spin pennies 40 times and get 17 heads. Should they reject the standard that pennies land heads 50% of the time?

**Solution**

Of course, Jenny and Maya shouldn’t expect to get exactly 20 heads (50% of 40) even if heads and tails are equally likely. There is variation in sampling—you don’t get exactly the same result each time you take a random sample from a population. Jenny and Maya construct the sampling distribution in Display 8.8, which shows the probabilities of getting different values of \( \hat{p} \) when \( n = 40 \) and \( p = 0.5 \).

**Display 8.8** Sampling distribution of \( \hat{p} \) for \( n = 40 \) and \( p = 0.5 \).
Jenny and Maya’s sample proportion, \( \hat{p} \), of \( \frac{17}{40} \) or 0.425 lies near the middle of this distribution. This means that if the percentage of heads actually is 50% when a penny is spun, getting 17 heads out of 40 spins is reasonably likely. So Jenny and Maya have no evidence to argue against the standard that a penny lands heads 50% of the time when spun.

A sample proportion is said to be \textbf{statistically significant} if it isn’t a reasonably likely outcome when the proposed standard is true. (The statement that the proposed standard is true is called the \textbf{null hypothesis}, indicated by the symbol \( H_0 \).) Jenny and Maya’s result of 17 heads out of 40 spins wasn’t statistically significant. Their value of \( \hat{p} \), 0.425, is a reasonably likely outcome if spinning a penny is fair.

\textbf{Example: Miguel and Kevin’s Spins}

Miguel and Kevin spin pennies and get 10 heads out of 40 spins for a sample proportion, \( \hat{p} \), of 0.25. Is this a statistically significant result?

\textbf{Solution}

As shown in Display 8.8, if 0.5 is the true proportion of heads when a penny is spun, it would be a rare event to get only 10 heads in a sample of 40 spins. Miguel and Kevin have a statistically significant result. This leads them to conclude that 0.5 probably is not the true proportion of heads when a penny is spun.

Another way of seeing this is to note that \( p = 0.5 \) isn’t included in the 95% confidence interval for the true proportion of heads when a penny is spun. Look at the vertical line at \( \hat{p} = 0.25 \) in Display 8.9. It doesn’t intersect the horizontal line that represents the population with \( p = 0.5 \). That is, the sample proportion \( \hat{p} = 0.25 \) is far enough from the standard of 0.5 to reject 0.5 as a plausible value for the proportion of times heads appear when a penny is spun.

\begin{displayquote}
\textbf{Rejection Region}
\end{displayquote}

A sample proportion of \( \hat{p} = 0.25 \) isn’t reasonably likely when \( p = 0.5 \).
When you spin a penny, there are three different proportions to keep straight:

- The true proportion of heads, \( p \), when a penny is spun. No one knows this value for sure. You might suspect from your own results that \( p \) does not equal 0.5.
- The proportion of heads, \( \hat{p} \), in your sample of 40 spins. Jenny and Maya’s sample proportion was \( \hat{p} = \frac{17}{40} \) or 0.425. Miguel and Kevin’s sample proportion was \( \hat{p} = \frac{10}{40} \) or 0.25.
- The proportion of heads that you hypothesize is the true proportion of heads when a coin is spun. The symbol used for this standard value is \( p_0 \). You have been testing the standard that spinning a penny is fair, so you have been using \( p_0 = 0.5 \).

**Informal Significance Testing**

D21. Consider the question of whether spinning a penny is fair. You will need your penny-spinning results from Activity 8.2a.

a. When investigating whether spinning a penny is fair, what is the standard?

b. What is your sample proportion, \( \hat{p} \), of heads from Activity 8.2a?

c. Locate your sample proportion, \( \hat{p} \), on the chart in Display 8.9. If spinning a penny is fair, is your sample proportion, \( \hat{p} \), reasonably likely?

d. What is your conclusion? Do your data provide evidence against the standard?

D22. Use your penny-flipping results from Activity 8.2a to answer the questions in D21.

**The Test Statistic**

Miguel and Kevin could see from Display 8.8 that it is not reasonably likely that they would get \( \hat{p} = 0.25 \) when \( p = 0.5 \), so they rejected the null hypothesis that a spun penny comes up heads half the time. How could they have determined this without constructing the entire distribution?

They need to find out if their value of \( \hat{p} \) lies in the middle 95% of all possible values of \( \hat{p} \). The distribution in Display 8.8 is approximately normal, so Miguel and Kevin can compute a \( z \)-score for their sample proportion, \( \hat{p} \), of 0.25:

\[
z = \frac{\text{statistic} - \text{parameter}}{\text{standard deviation of statistic}}
\]

\[
= \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}
\]

\[
= \frac{0.25 - 0.5}{\sqrt{\frac{0.5(1 - 0.5)}{40}}}
\]

\[
\approx -3.16
\]
Having a sample proportion that is 3.16 standard errors below the hypothesized mean \( p_0 \) of 0.5, would definitely be a rare event if it is true that half of spun pennies land heads. The value of \( \hat{p} \) is outside the reasonably likely outcomes for \( p = 0.5 \), so Miguel and Kevin reject the null hypothesis that spinning a penny is fair. The \( z \)-score that Miguel and Kevin computed is an example of a test statistic.

### The Test Statistic for Testing a Proportion

To determine if \( \hat{p} \) is reasonably likely or a rare event for a given standard \( p_0 \), you need to check the value of the test statistic

\[
z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}
\]

This tells you how many standard errors the sample proportion \( \hat{p} \) lies from the standard, \( p_0 \).

### DISCUSSION

D23. What does it mean if the test statistic has the value 0?

D24. If everything else remains constant, what will happen to the test statistic \( z \) when

- the sample size \( n \) increases?
- \( \hat{p} \) gets farther from \( p_0 \)?

D25. Could Miguel and Kevin be wrong when they decide that spinning a penny isn’t fair? Explain.

### P-Values

Instead of simply reporting whether a result is statistically significant, it is common practice also to report a \( P \)-value.

The **P-value** for a test is the probability of seeing a result from a random sample that is as extreme as or more extreme than the result you got from your random sample if the null hypothesis is true.

In Chapter 1 you found the equivalent of a \( P \)-value by simulation when examining the chance that the mean age of three randomly selected laid-off workers would exceed a given value. Now, you will learn a way to find such values by theory rather than by simulation.

### Example: Interpreting a \( P \)-Value

Chimpanzees were given the choice of pulling one of two heavy rakes toward them. Pulling one of the rakes would bring food with it. Pulling the other rake would not bring in food. Display 8.10 shows a typical setup for this study. The round object represents the food. The researchers found that the correct rake was
chosen significantly more often than would be expected by chance ($P = 0.01$).
Interpret this $P$-value in the context of the situation. [Source: Victoria Horner and Andrew Whiten, “Causal Knowledge and Imitation/Emulation Switching in Chimpanzees (Pan troglodytes) and Children (Homo sapiens),” *Animal Cognition* 8 (2005): 164–81.]

**Display 8.10** Diagram of chimpanzee rake setup.

**Solution**
The $P$-value 0.01 means that if the chimps were selecting a rake at random, which gives them a 50–50 chance of selecting the one that rakes in the food, the probability is only 0.01 that the chimps would rake in the food as often as or more often than they did. The researchers therefore can conclude that the chimps were able to deliberately choose the rake that got them the food.

**Example: Computing a $P$-Value for Jenny and Maya**
Find the $P$-value for Jenny and Maya’s test of whether spinning a penny is fair.

**Solution**
Jenny and Maya got 17 heads out of 40 spins, so the value of the test statistic is

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} = \frac{0.425 - 0.5}{\sqrt{\frac{0.5(1 - 0.5)}{40}}} \approx -0.95$$

Assume for now that the null hypothesis is true: Spinning a penny gives you heads 50% of the time. The probability of getting a value of $z$ less than $-0.95$ can be found in Table A on page 824 or from a calculator. This value is approximately 0.171, as shown in Display 8.11.

**Display 8.11** The probability of 17 or fewer heads with 40 spins of a coin, if spinning a coin is fair.

However, Jenny and Maya notice that getting 23 or more heads is just as extreme as getting 17 heads or fewer. Now the sketch looks like Display 8.12.
The probability of getting a result as far out in the tails of the distribution as they did is $2(0.171)$, or 0.342. This is their $P$-value. If the probability of getting heads is 0.5, there is a 34.2% chance of getting 17 heads or fewer or getting 23 heads or more in 40 spins. Because this $P$-value is relatively large, Jenny and Maya don’t have statistically significant evidence to conclude that spinning a penny is anything other than fair.

Display 8.12  The probability of at most 17 or at least 23 heads with 40 spins of a coin, if spinning a coin is fair.

[See Calculator Note 8D to learn how to find a value of $z$ using your calculator.]

In summary, the test statistic, $z$, is computed using the hypothesized value, $p_0$, as the mean. A small $P$-value tells you that the sample proportion you observed is quite far away from $p_0$. Your data aren’t behaving in a manner consistent with the null hypothesis. A large $P$-value tells you that the sample proportion you observed is near $p_0$, so your result isn’t statistically significant. So, the $P$-value weighs the evidence found from the data: A small $P$-value places the weight against the null hypothesis, and a large $P$-value weighs in as consistent with the null hypothesis.

**P-Values**

D26. Why is the phrase “if the null hypothesis is true” necessary in the interpretation of a $P$-value?

D27. Does the $P$-value give the probability that the null hypothesis is true?

**Critical Values and Level of Significance**

You have been rejecting the hypothesized standard when $\hat{p}$ would be a rare event if that standard were true. This is equivalent to rejecting the standard when the value of the test statistic, $z$, is in the outer 5% of the standard normal distribution, that is, when $z$ is less than $-1.96$ or greater than $1.96$. Dividing points such as $\pm 1.96$ are called critical values (denoted $\pm z^*$). The corresponding proportion, 0.05 in this case, is called the level of significance (denoted $\alpha$). Other critical values can be used depending on the level of significance desired.
Example: Finding Critical Values

Suppose you want to reject the null hypothesis when the test statistic, $z$, is in the outer 10% of the standard normal distribution—that is, with a level of significance equal to 0.10. What should you use as critical values? Should you reject the null hypothesis if $z$ turns out to be 1.87?

Solution

Find the values of $z$ in Table A on page 824 that cut off the top 5% and the bottom 5% of the standard normal distribution. The value that cuts off the bottom 5% lies about halfway between $-1.64$ and $-1.65$. Thus, the critical values for $\alpha = 0.10$ are $\pm 1.645$, as shown in Display 8.13. (For most critical values, you can simply use the nearest value of $z$, to two decimal places.) You can also use your calculator to find critical values. [See Calculator Note 8B.]

![Critical Values](image)

Display 8.13: The test statistic $z = 1.87$ is more extreme than $z^* = 1.645$.

Because $z = 1.87$ is more extreme than $z^* = 1.645$, reject the null hypothesis. The sample proportion, $\hat{p}$, is farther from $p_0$ than would be reasonably likely if $p_0$ were the true population proportion.

The box summarizes the use of critical values.

Using Critical Values and the Level of Significance

If the value of the test statistic $z$ is more extreme than the critical values, $\pm z^*$, you have chosen (or, equivalently, the $P$-value is less than $\alpha$), you have evidence against the null hypothesis. Reject the null hypothesis and say that the result is statistically significant.

On the other hand, if $z$ is less extreme than the critical values (or, equivalently, the $P$-value is greater than $\alpha$), you do not have evidence against the null hypothesis and so you cannot reject it.

If a level of significance isn’t specified, it is usually safe to assume that $\alpha = 0.05$ and $z^* = \pm 1.96$. 
Critical Values and Level of Significance

D28. Suppose your null hypothesis is that spinning (or flipping) a coin is fair. Find the test statistic, $z$, for your results from Activity 8.2a. Using $\alpha = 0.05$, give your conclusion
   a. for spinning 40 pennies
   b. for flipping 40 pennies

D29. If everything else remains constant, is it easier to reject the null hypothesis when the level of significance, $\alpha$, is larger or smaller?

D30. Select the correct words in each sentence.
   a. A larger critical value makes the level of significance, $\alpha$, smaller/larger and so it will be harder/easier to reject the null hypothesis.
   b. A smaller critical value makes the level of significance, $\alpha$, smaller/larger and so it will be harder/easier to reject the null hypothesis.

The Formal Language of Tests of Significance

You have been using informal language to describe tests of significance. More formal terminology appears in the box. You will examine each component of a significance test thoroughly in the rest of this section.

Components of a Significance Test for a Proportion

1. **Give the name of the test and check the conditions for its use.** For a significance test for a proportion, three conditions must be met:
   - The sample is a simple random sample from a binomial population.
   - Both $np_0$ and $n(1 - p_0)$ are at least 10.
   - The population size is at least 10 times the sample size.

2. **State the hypotheses, defining any symbols.** When testing a proportion, the null hypothesis, $H_0$, is
   
   $H_0$: The percentage of successes, $p$, in the population from which the sample came is equal to $p_0$.

   The alternative hypothesis, $H_a$, can be of three forms:
   
   $H_a$: The percentage of successes, $p$, in the population from which the sample came is not equal to $p_0$.
   
   $H_a$: The percentage of successes, $p$, in the population from which the sample came is greater than $p_0$.
   
   $H_a$: The percentage of successes, $p$, in the population from which the sample came is less than $p_0$.

(continued)
3. **Compute the test statistic, \( z \), and find the critical values, \( z^* \), and the \( P \)-value.** Include a sketch that illustrates the situation.

The test statistic is

\[
z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}
\]

4. **Write a conclusion.** There are two parts to stating a conclusion:

- Compare the value of \( z \) to the predetermined critical values, or compare the \( P \)-value to \( \alpha \). Then say whether you reject the null hypothesis or don't reject the null hypothesis, linking your reason to the \( P \)-value or to the critical values.
- Tell what your conclusion means in the context of the situation.

---

**Example: Jenny and Maya**

Illustrate the components of a significance test using Jenny and Maya’s results from spinning a penny 40 times and getting 17 heads. Use \( \alpha = 0.05 \).

**Solution**

1. Apparently, Jenny and Maya spun the pennies carefully and independently, so we can consider the number of heads they got to be a binomial random variable. Both \( np_0 \) and \( n(1 - p_0) \) are at least 10, because \( np_0 = 40(0.5) \), or 20, and \( n(1 - p_0) = 40(0.5) \), or 20. The population of all possible spins is infinitely large, so it definitely is at least 10 times their sample size of 40. The conditions for doing a significance test for a proportion are met.

2. \( H_0: \) The proportion of heads when a penny is spun is equal to 0.5.

   \( H_a: \) The proportion of heads when a penny is spun is not equal to 0.5.

   Or you could write

   \( H_0: \) \( p = 0.5 \), where \( p \) is the proportion of heads when a penny is spun

   \( H_a: \) \( p \neq 0.5 \)

3. As you saw in the example on page 495, the test statistic is

\[
z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}
\]

\[
= \frac{0.425 - 0.5}{\sqrt{\frac{0.5(1 - 0.5)}{40}}}
\]

\[
\approx -0.95
\]
The critical values are $\pm 1.96$ and the $P$-value is 0.342. A sketch illustrating the situation is shown in Display 8.14.

4. Because $z = -0.95$ is less extreme than the critical value, $-1.96$, this isn’t a statistically significant result. Equivalently, because the $P$-value, 0.342, is larger than $\alpha = 0.05$, this isn’t a statistically significant result.

If spinning a penny results in heads 50% of the time, you are reasonably likely to get only 17 heads out of 40 flips. Thus, Jenny and Maya cannot reject the null hypothesis that spinning a penny is a fair process. (You should never say that you “accept” the null hypothesis, because if you constructed a confidence interval for $p$, all the other values in it—not just 0.5—are also plausible values of $p$.)

You will more easily remember the structure of a test of significance if you keep in mind what you should do before you look at the data from the sample.

- **Check the conditions for the test.** Checking the conditions to be met for a significance test for a proportion requires only knowledge of how the sample was collected, the sample size, and the value of the hypothesized standard, $p_0$. (For some tests, as you will learn later, you will have to peek at the data in order to check conditions.)

- **Write your hypotheses.** The hypotheses should be based on the research question to be investigated, not on the data. Ideally, the investigator sets the hypotheses before the data are collected. This means that the value of $\hat{p}$ should not appear in your hypotheses.

- **Decide on the level of significance.** The level of significance, $\alpha$, and the corresponding critical values, $\pm z^*$, are set by the investigator before the data are collected. Some fields have set standards for the level of significance, typically $\alpha = 0.05$.

- **Sketch the standard normal distribution.** Mark the critical values, $\pm z^*$. You can place the value of the test statistic, $z$, on this distribution after you look at the data.

You will use the data (number of successes and value of $\hat{p}$) only to compute the test statistic and to write your conclusion.
D31. What is the difference in the meaning of the symbols $p$, $p_0$, and $\hat{p}$? Of the three, which varies depending on the sample you select? Which of the three is unknown?

D32. A large study of births in the United States found that about 51.2% of all babies born in the United States are boys and about 48.8% are girls. This is a statistically significant difference, meaning that the difference can't reasonably be attributed to chance. [Source: Centers for Disease Control, 2005, www.cdc.gov; Trend Analysis of the Sex Ratio at Birth in the United States, National Vital Statistics Reports 53, no. 20 (June 14, 2005): 1.]

a. Is this result of any practical significance to a pair of prospective parents who were hoping for a girl?

b. Give an example where this difference would have some practical significance.

D33. Suppose the null hypothesis is true and you have a random sample of size $n$.

a. What is the approximate probability that the test statistic will be greater than 1.96? What is the probability that it will either be larger than 1.96 or be less than $-1.96$?

b. What is the probability that the test statistic will exceed 1.645? What is the probability that it will either be larger than 1.645 or be less than $-1.645$?

c. Why do you need to know that $H_0$ is true to answer parts a and b?

Types of Errors

The reasoning of significance tests often is compared to that of a jury trial. The possibilities in such a trial are given in this diagram.

<table>
<thead>
<tr>
<th>Defendant Is Actually</th>
<th>Not Guilty</th>
<th>Guilty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Innocent</td>
<td>Correct</td>
<td>Error</td>
</tr>
<tr>
<td>Guilty</td>
<td>Worse error</td>
<td>Correct</td>
</tr>
</tbody>
</table>

In the same way, there are two types of errors in significance testing.

<table>
<thead>
<tr>
<th>Null Hypothesis Is Actually</th>
<th>Don’t Reject $H_0$</th>
<th>Reject $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>Correct</td>
<td>Type I error</td>
</tr>
<tr>
<td>False</td>
<td>Type II error</td>
<td>Correct</td>
</tr>
</tbody>
</table>

Miguel and Kevin got a test statistic with a large absolute value and so concluded that spinning a penny is not fair. However, Jenny and Maya got a test statistic with a value that was close to 0, so the result from their sample was quite consistent with the idea that spinning a penny is fair. Who is right? Thousands of
statistics students have contributed to the solution of this problem by spinning pennies. Their instructors have reported that the true proportion of heads seems to be around 0.4. So Miguel and Kevin made the correct decision. Jenny and Maya have made the error of not rejecting a null hypothesis that is false—they have made a Type II error.

If, like Miguel and Kevin, your test statistic is large in absolute value, then there are several possibilities to consider:

- The null hypothesis is true and a rare event occurred. That is, it was just bad luck that resulted in \( \hat{p} \) being so far from \( p_0 \).
- The null hypothesis isn’t true, and that’s why the sample proportion, \( \hat{p} \), was so far from \( p_0 \).
- The sampling process was biased in some way, so the value of \( \hat{p} \) is itself suspicious.

If you can rule out the last possibility, then the usual decision is to reject the null hypothesis. However, you might be making a Type I error—rejecting \( H_0 \) even though \( H_0 \) is actually true.

If, like Jenny and Maya, your test statistic is small in absolute value, then there are also several possibilities to consider:

- The null hypothesis is true, and you got just about what you would expect in the sample.
- The null hypothesis isn’t true, and it was just by chance that \( \hat{p} \) turned out to be close to \( p_0 \).
- The sampling process was biased in some way, so the value of \( \hat{p} \) is itself suspicious.

If you can rule out the last possibility, then the usual decision is to say that you cannot reject the null hypothesis. However, you might be making a Type II error, as Jenny and Maya did. This doesn’t mean they did anything wrong. The error occurred because the results from their sample just happened to be consistent with the null hypothesis.

**Example: An ESP Test and a Type I Error**

Several Internet sites allow you to take a test to see if you have extrasensory perception (ESP). Ann took one of these tests. She was presented with the back of one of the five different cards described in E22 in Chapter 5 (page 314). She guessed which of the five cards it was. In 100 tries with different cards, she guessed right 29 times. The Web site said that her result was “significant” and the evidence for ESP was “fair.” What is the \( P \)-value for this test? Discuss what type of error might have been made.

**Solution**

Conditions have been met to use a normal approximation. The test statistic is

\[
z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} = \frac{0.29 - 0.20}{\sqrt{\frac{0.20(1 - 0.20)}{100}}} = 2.25
\]
The corresponding $P$-value is 0.02, as illustrated in Display 8.15.

![Display 8.15](image)

**Display 8.15** The $P$-value for a $z$-score of 2.25.

Ann knows that she does not have ESP and, in fact, didn’t even try to discern what the card was, instead selecting her choices rapidly and at random. In other words, the null hypothesis is true that, in the long run, she will guess the correct card 20% of the time. A Type I error has been made. Out of every 100 people who take such a test, we expect that two of them will get 29 or more cards right (or 11 or fewer cards right). Ann was one of the lucky two. When she tried the ESP test again, she correctly identified only 17 cards.

Suppose the null hypothesis is true. What is the chance that you will reject it, making a Type I error? The only way you can make a Type I error is to get a rare event from your sample. For example, if you are using a significance level of 0.05, you would make a Type I error if you get a value of the test statistic larger than 1.96 or smaller than $-1.96$. This happens only 5% of the time no matter what the sample size. Thus, if the null hypothesis is true, the probability of a Type I error is 0.05. If you used $z^* = \pm 2.576$, the critical values for a significance level of 0.01, then the probability of making a Type I error would be only 0.01. If the null hypothesis is true, the probability of a Type I error is equal to the level of significance. To lower the chance of a Type I error, then, your best strategy is to have a low level of significance or, equivalently, large critical values.

### DISCUSSION

**Types of Errors**

D34. Suppose you use critical values of $\pm 2$ and the null hypothesis is actually true. What is the probability that you will get a sample that results in rejecting the null hypothesis? Explain. What type of error is this?

D35. To avoid Type I errors, why not always use a very large critical value?

**Power and Type II Errors**

If Jenny and Maya had spun more pennies, they soon would have realized that the true proportion of heads is less than 0.5. However, their sample size was too small for their test to detect a significant difference between the proportion of heads in their sample and the hypothesized value, 0.5. So their analysis resulted in a Type II error, that is, a failure to reject a false null hypothesis. Another way to state this is to say that their test did not have enough power to be able to detect that the true proportion of heads, $p$, is different from the proportion hypothesized, $p_0$. 

8.2 Testing a Proportion 503
The probability of a Type II error is denoted \( \beta \).

**Power of a Test**

The **power** of a test is the probability of rejecting the null hypothesis. When the null hypothesis is false,

\[
power = 1 - \text{probability of a Type II error}
\]

It follows that if the probability of a Type II error is small, then the power of the test is large.

**Example: Jenny and Maya Get More Power**

Jenny and Maya start again and this time spin 200 pennies instead of only 40. They happen to get the same proportion of heads, \( \hat{p} = 0.425 \), that they got with only 40 spins. If they use \( \alpha = 0.05 \) as the level of significance, what conclusion would they come to now? Explain why this is the case.

**Solution**

Their new test statistic is

\[
z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} = \frac{0.425 - 0.5}{\sqrt{\frac{0.5(1 - 0.5)}{200}}} = -2.12
\]

Because \( z = -2.12 \) is more extreme than the critical value, \( -1.96 \), Jenny and Maya now can come to the correct conclusion that the proportion of heads when a penny is spun is not 0.5.

The reason why the larger sample size gave a statistically significant result and the smaller sample size did not is illustrated in the diagrams that follow. Display 8.16 shows that a sample proportion, \( \hat{p} \), of 0.425 is reasonably likely for a fair coin spun 40 times. With 40 spins, Jenny and Maya have no evidence to reject the hypothesis that spinning is fair.

![Display 8.16](image)
Display 8.17 shows that a sample proportion, \( \hat{p} \), of 0.425 is not reasonably likely for a fair coin spun 200 times. The difference between the two distributions is their spread. The spread is controlled by the sample size: bigger sample, less spread. If spinning a coin is fair, Jenny and Maya should get quite close to 50% heads with 200 spins. Because they were far away from 0.50 in this distribution, they can correctly conclude that spinning a coin doesn’t result in 50% heads. In other words, the larger sample size produces greater power to make the correct decision.

As the previous example shows, if the null hypothesis is false and should be rejected, a larger sample size increases the probability that you will be able to reject it. However, if the null hypothesis is true, increasing the sample size has no effect on the probability of making a Type I error. The only way you can decrease the probability of making a Type I error is to make the level of significance, \( \alpha \), smaller. As you’ll see in the next example, however, if the null hypothesis is actually false, this strategy results in a higher probability of a Type II error and lower power!

**Example: Miguel and Kevin Lose Power and Make an Error**

Suppose Miguel and Kevin were worried about making a Type I error. Consequently, they decided to use a 0.001 level of significance. That is, if it is true that spinning a penny is fair, they wanted to have only a 0.001 chance of deciding wrongly that spinning isn’t fair. What is their conclusion now?

**Solution**

With \( \alpha = 0.001 \), the critical values found from a calculator are \( \pm 3.29 \). (If you are using Table A on page 824, you will find that \(-3.27\) cuts off an area of approximately 0.0005.) With this level of significance, the value of their test statistic, \( z = -3.16 \), is inside the critical values, so Miguel and Kevin would not have rejected the false null hypothesis that spinning a penny is fair. They would have made a Type II error because their test lost power.
In the previous examples, you observed three important properties of power:

- Power increases as the sample size increases, all else being held constant.
- Power decreases as the value of $\alpha$ decreases, all else being held constant.
- Power increases when the true population proportion, $p$, is farther from the hypothesized value, $p_0$. When $p$ is farther from $p_0$, $\hat{p}$ tends to be farther from $p_0$, so the test statistic tends to be larger.

The box summarizes what you should understand about types of error and power.

### Types of Error and Power

**Type I Error**

When the null hypothesis is true and you reject it, you have made a Type I error. The probability of making a Type I error is equal to the significance level, $\alpha$, of the test. To decrease the probability of a Type I error, make $\alpha$ smaller. Changing the sample size has no effect on the probability of a Type I error.

If the null hypothesis is false, you can’t make a Type I error.

**Type II Error**

When the null hypothesis is false and you fail to reject it, you have made a Type II error. To decrease the probability of making a Type II error, take a larger sample or make the significance level, $\alpha$, larger.

If the null hypothesis is true, you can’t make a Type II error.

**Power**

Power is the probability of rejecting the null hypothesis.

When the null hypothesis is false, you want to reject it and therefore you want the power to be large. To increase power, you can either take a larger sample or make $\alpha$ larger.

### Discussion

**Power and Type II Errors**

D36. Explain why an increase in sample size increases the power of a test, all else remaining unchanged.

D37. What happens to the power of a test as the population proportion, $p$, moves farther away from the hypothesized value, $p_0$, all else remaining unchanged?

D38. Recall that the power of a test is the probability of rejecting the null hypothesis. Can a statistical test of the type discussed in this chapter ever have a power of 1? Can a statistical test of the type discussed in this chapter ever have a power of 0? If so, would either be desirable from a practical point of view?
**One-Sided Tests of Significance**

When testing the effectiveness of a new drug, the investigator must establish that the new drug has a better cure rate than the older treatment (or that there are fewer side effects). He or she isn’t interested in simply rejecting the null hypothesis that the new drug has the same cure rate as the older treatment but instead needs to know if it is better. In such situations, the alternative hypothesis should state that the new drug cures a larger proportion of people than does the older treatment. This is called a one-sided test of significance (or one-tailed test of significance). Tests of significance can be one-sided if the investigator has an indication of which way any change from the standard should go. This must be decided before looking at the data.

The tests of significance discussed previously were two-sided—the investigator was interested in detecting a change from the standard in either direction. For example, Jenny and Maya were testing to see if the percentage of heads when a penny is spun is different from 0.5.

In testing a proportion, the alternative hypothesis can take one of three forms:

- **H₁:** The percentage of successes \( p \) in the population from which the sample came is not \( p_0 \). In symbols, \( p \neq p_0 \).

- **H₁:** The percentage of successes \( p \) in the population from which the sample came is greater than \( p_0 \). In symbols, \( p > p_0 \).

- **H₁:** The percentage of successes \( p \) in the population from which the sample came is less than \( p_0 \). In symbols, \( p < p_0 \).

The first alternative hypothesis is for a two-sided test; the latter two are for one-sided tests. For a one-sided test, the \( P \)-value is found using one side only.

**Example: One-Sided Test of Significance**

The editors of a magazine have noticed that people seem to believe that a successful life depends on having good friends. They would like to have a story about this and use a headline such as “Most Adults Believe Friends Are Important for Success.” So they have commissioned a survey to ask a random sample of adults whether a successful life depends on having good friends. In a random sample of 1027 adults, 53% said yes. Should the editors go ahead and use their headline?

**Solution**

As always, the answer must include statistical evidence, in this case a test of significance. You will use \( p_0 = 0.5 \) because “most” means greater than 0.5.
1. The sample was a simple random sample of adults in the United States. Both \( np_0 = 1027(0.5) = 513.5 \) and \( n(1 - p_0) = 1027(0.5) = 513.5 \) are at least 10. The population of all adults in the United States is much larger than 10(1027) = 10,270.

2. The claim (the alternative hypothesis) is that “most” adults believe this. This is the same as “more than 50%.” Let \( p \) represent the proportion of adults in the United States who believe that a successful life depends on having good friends. The null hypothesis is that the percentage who believe this is 50% (or less):

\[
H_0: \ p = 0.5
\]

The border-line equality, \( p = 0.5 \), can be used for the null hypothesis because rejection of this value in favor of some value larger than 0.5 would certainly imply that any smaller value of \( p \) would be rejected as well.

The alternative hypothesis is that this percentage is more than 50%:

\[
H_a: \ p > 0.5
\]

3. The test statistic is

\[
z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} = \frac{0.53 - 0.5}{\sqrt{0.5(1 - 0.5)/1027}} \approx 1.92
\]

As shown in Display 8.18, the \( P \)-value of 0.0274 is the probability of getting a test statistic greater than 1.92 if the null hypothesis is true.

4. The \( P \)-value (found in the upper tail) is fairly small. If the percentage of all adults who believe a successful life depends on having good friends is 50%, then the probability of getting a sample proportion of 53% or more is only 0.0274. Because getting a sample proportion, \( \hat{p} \), of 0.53 or more is so unlikely if the null hypothesis is true, reject the null hypothesis. This is quite strong evidence that the true percentage must be greater than 50%. The editors should feel free to use their headline.

**DISCUSSION**

**One-Sided Tests of Significance**

D39. What would have been the conclusion if 40% of the sample in the preceding example believed that friends are important for success?

D40. If everything else remains the same, is it easier to reject a false null hypothesis with a one-sided test or with a two-sided test?
Summary 8.2: Tests of Significance

Suppose you want to decide whether it is reasonable to assume that your random sample comes from a population with proportion of successes $p_0$. A test of significance for a proportion tells you whether the results from your sample are so different from what you would expect that you should reject that value of $p_0$ as a plausible value. If the sample proportion, $\hat{p}$, is relatively close to $p_0$, you can reasonably attribute the difference to chance variation and should not reject $p_0$ as the possible proportion of successes in the population. If the sample proportion, $\hat{p}$, is relatively far from $p_0$, you cannot reasonably attribute the difference to chance variation and should reject $p_0$ as the possible proportion of successes in the population.

The four components of a test of significance are listed here, with some suggestions about how to interpret each one.

- **Checking conditions.** In real surveys, it is almost always the case that the sample was not a simple random sample from the population. If the sample comes from a more complicated design that uses randomization, inference is possible. In other cases, convenience samples are used—people are chosen for a sample because they were easy to find. In such cases, confidence intervals and significance tests generally should not be used.

- **Writing hypotheses.** Remember that, when you write the hypotheses, you are not supposed to have seen the sample yet. Thus, the value from the sample, $\hat{p}$, does not appear in the null hypothesis. If you use the symbol $p$ in the null hypothesis, always say in words what it represents in the context of the problem.

- **Doing computations.** The test statistic measures how compatible the result from the sample is with the null hypothesis. Test statistics typically come in this form:

  $$\frac{\text{statistic} - \text{parameter}}{\text{standard deviation of statistic}}$$

  In the case of testing a proportion, this becomes

  $$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

- **Writing a conclusion.** If $|z|$ is large and thus the $P$-value is small (smaller than some significance level $\alpha$ that you have decided on in advance), this is evidence against the null hypothesis and so you should reject it. If the $P$-value is larger than the predetermined value $\alpha$, do not reject the null hypothesis.
If you reject the null hypothesis, you are saying that the difference between \( \hat{p} \) and \( p_0 \) is so large that you can’t reasonably attribute it to chance variation. If you don’t reject the null hypothesis, you are saying that the difference between \( \hat{p} \) and \( p_0 \) is small enough so that it looks like the kind of variation you would expect from a binomial situation in which the proportion of successes is \( p_0 \).

Always write your conclusion in the context of the situation. Do not simply write “Reject the null hypothesis” or “Do not reject the null hypothesis.” Explain what your conclusion means in terms of the context of the problem.

You should never write “Accept the null hypothesis.” The reason is that the sample can tell you only whether the null hypothesis is reasonable. If you get a very small \( P \)-value, then it isn’t reasonable to assume that the null hypothesis is true and you should reject it. However, if you get a large \( P \)-value, then the null hypothesis is plausible but still might not give the exact value of \( p \).

Finally, just because a result is statistically significant doesn’t mean the difference has any practical significance in the real world. In other words, the difference might exist, but it might be so small as to have no impact.

**Practice**

**Informal Significance Testing**

P20. A 1997 article reported that two-thirds of teens in grades 7–12 want to study more about medical research. You wonder if this proportion still holds today and decide to test it. You take a random sample of 40 teens and find that only 23 want to study more about medical research. [Source: CNN Interactive Story Page, April 22, 1997, www.cnn.com.]

a. What is the standard (the hypothesized value, \( p_0 \), of the population proportion)?

b. What is the sample proportion, \( \hat{p} \)?

c. Use Display 8.9 to determine whether the result is statistically significant. That is, is there evidence leading you to believe that the proportion today is different from the proportion in 1997?

P21. A student took a 40-question true–false test and got 30 answers correct. The student says, “That proves I was not guessing at the answers.”

a. What is the standard, \( p_0 \)?

b. What is the sample proportion, \( \hat{p} \)?

c. Use Display 8.9 to determine whether the result is statistically significant.

d. Does the answer to part c prove that the student was not guessing?

P22. This year, 75% of the seniors wanted extra tickets for their graduation ceremony. To anticipate whether there might be a change in that percentage next year, the junior class took a random sample of 40 juniors and found that 32 would want extra tickets.

a. What is the standard, \( p_0 \)?

b. What is the sample proportion, \( \hat{p} \)?

c. Use Display 8.9 to determine whether this result is statistically significant.

d. Is this statistical evidence of a change?

**The Test Statistic**

P23. For the situation in P20, what value of the test statistic should the junior class use to test whether there is statistical evidence of a change?

P24. Forty-five dogs and their owners, chosen at random, were photographed separately. A judge was shown a picture of each owner and pictures of two dogs and asked to pick the dog that went with the owner. The judge was right 23 times. What value of the test statistic should be used to test whether the judge did better than could reasonably be expected just by guessing? [Source: Based on a study published in Psychological Science, 2004.]
**P-Values**

P25. Find the $P$-value for Miguel and Kevin’s test statistic, $-3.16$, from pages 493–494. Write a sentence explaining what this $P$-value means in the context of their situation.

P26. Suppose you re-create the chimpanzee experiment in the example on page 494 with an apparatus that tests whether cats can select the rake that results in the food. You test 50 randomly chosen cats and find that 28 select the rake that pulls in the food. Find the $P$-value for this test and interpret it.

P27. According to the U.S. Census Bureau, about 69% of houses across the country are occupied by their owners. Your class randomly samples 50 houses in your community and finds that 30 houses are occupied by the owners. Follow these steps to determine whether your community differs from the nation as a whole as to the percentage of houses that are owner-occupied.
   a. State the null hypothesis to be tested. What is the alternative hypothesis?
   b. Find the value of the test statistic.
   c. Find the $P$-value for this test and explain what it means.
   d. Write a conclusion, based on your analysis, in the context of the problem.

P28. Which of statements A–E is the best explanation of what is meant by the $P$-value of a test of significance?
   A. Assuming that you had a random sample and the other conditions for a significance test are met, the $P$-value is the probability that $H_0$ is true.
   B. Assuming that $H_0$ is true, the $P$-value is the probability of observing a value of a test statistic at least as far out in the tails of the sampling distribution as is the value of $z$ from your sample.
   C. The $P$-value is the probability that $H_0$ is false.
   D. Assuming that the sampling distribution is normal, the $P$-value is the probability that $H_0$ is true.
   E. Assuming that $H_0$ is true, the $P$-value is the probability of observing the same value of $z$ that you got in your sample.

**Critical Values and Level of Significance**

P29. If your null hypothesis is that spinning or flipping a coin is fair, find the test statistic, $z$, for each of these situations and give your conclusion. Use $\alpha = 0.05$.
   a. A class spun quarters 500 times and got 194 heads.
   b. A class flipped quarters 500 times and got 265 heads.

P30. Use Table A on page 824 or your calculator to answer these questions.
   a. What critical value is associated with a significance level of 0.12?
   b. What significance level is associated with critical values of $z^*$ of $\pm 1.73$?

P31. Use Table A on page 824 or your calculator to answer these questions.
   a. What level of significance is associated with critical values of $z^*$ of $\pm 2.576$?
   b. What critical values are associated with a significance level of 2%?

**The Formal Language of Tests of Significance**

P32. Using the data in P29 that there were 194 heads out of 500 spins, carry out the four steps in a test of the null hypothesis that spinning a quarter is fair.

P33. Suppose that, in a random sample of 500 bookstores across the United States, 265 also sell DVDs. Carry out the four steps in a test of the null hypothesis that half the bookstores in the United States sell DVDs.

P34. State an appropriate null hypothesis for each of these situations.
   a. You wonder if a student is guessing on a multiple-choice test of 60 questions where each question has five possible answers.
   b. You wonder if there is an equal proportion of male and female newscasters on local television stations.
c. You wonder if people who wash their cars once a week are most likely to wash them on Saturday.

P35. The United States 2000 Census found that 4% of all households were multigenerational (consisting of three or more generations of parents and their children). You want to test the null hypothesis that this percentage is the same this year as it was in 2000. You write the null hypothesis as \( p = 0.04 \). Which of these statements best describes what \( p \) stands for? [Source: www.census.gov.]

A. the proportion of all households that were multigenerational in 2000
B. the proportion of all households that are multigenerational this year
C. the proportion of multigenerational households in the sample in 2000
D. the proportion of multigenerational households in the sample this year
E. the proportion of all multigenerational households in both 2000 and this year

Types of Errors

P36. Hila is rolling a pair of dice to test whether they land doubles \( \frac{1}{6} \) of the time. She doesn’t know it, but the dice are actually fair. She will use a significance level, \( \alpha \), of 0.05. Hila rolls the dice 100 times and gets doubles 22 times.

a. What conclusion should Hila come to?
b. Did Hila make an error? If so, which type?

P37. Jeffrey and Taline each want to test if the proportion of adults in their neighborhood who have graduated from high school is 0.94, as claimed in the newspaper. Jeffrey takes a random sample of 200 adults and uses \( \alpha = 0.05 \). Taline takes a random sample of 500 adults and uses \( \alpha = 0.05 \). Suppose the newspaper’s percentage is actually right.

a. Is it possible for Jeffrey or Taline to make a Type I error? If so, who is more likely to do so?
b. Is it possible for Jeffrey or Taline to make a Type II error? If so, who is more likely to do so?

P38. Jeffrey and Taline each want to test if the proportion of adults in their neighborhood who took chemistry in high school is 0.25, as claimed in the newspaper. Jeffrey takes a random sample of 200 adults and uses \( \alpha = 0.05 \). Taline takes a random sample of 500 adults and uses \( \alpha = 0.05 \). Suppose the newspaper’s percentage is actually wrong.

a. Is it possible for Jeffrey or Taline to make a Type I error? If so, who is more likely to do so?
b. Is it possible for Jeffrey or Taline to make a Type II error? If so, who is more likely to do so?

Power and Type II Errors

P39. A statistical test is designed with a significance level of 0.05 and a sample size of 100. A similar test of the same null hypothesis is designed with a significance level of 0.10 and a sample size of 100. If the null hypothesis is false, which test has the greater power? Sketch a diagram to support your answer.

P40. A statistical test is designed with a significance level of 0.05 and a sample size of 100. A similar test of the same null hypothesis is designed with a significance level of 0.05 and a sample size of 200. If the null hypothesis is false, which test has the greater power? Sketch a diagram to support your answer.
P41. Suppose a medical researcher wants to test the hypothesis that the proportion of patients who develop undesirable side effects from a certain medication is the reported 6%. The researcher is pretty sure that the true proportion is close to either 8% or 10%. She has a random sample of patients using the medication and chooses a 5% level of significance. Is the power of the test larger if the true proportion is 8% or 10%? Sketch a diagram to support your answer.

One-Sided Tests of Significance

P42. In a poll of 1000 randomly sampled adults in the United States, 46% said they were satisfied with the quality of K–12 education in the nation. Does this imply that less than a majority of adult residents are satisfied with the quality of education? Answer this question by working through the steps in parts a–d. [Source: Gallup, Slim Majority Dissatisfied with Education in the U.S., 2005, poll.gallup.com.]

a. Verify that the conditions for the test are satisfied.

b. What is the null hypothesis? What is the most appropriate alternative hypothesis? Explain your reasoning.

c. Calculate a test statistic and the corresponding P-value.

d. Write a conclusion in the context of the original question.

P43. In the same poll of adults as described in P42, 51% of the 1000 respondents said that they were dissatisfied with the quality of K–12 education in schools today. Does this imply that a majority of adult residents are dissatisfied with the quality of education? Work through the same steps as in P42.

P44. You claim that the percentage of teens in your community who know that epilepsy is not contagious is larger than the national percentage, 51%. (See E5 on page 486.) You take a random sample of 169 teens in your community and find that 55% know that epilepsy is not contagious. Carry out the steps of a test of your claim.

P45. Determine whether each statement is true or false.

a. Always look at the data before writing the null and alternative hypotheses.

b. All else being equal, using a one-sided test will result in a larger P-value than using a two-sided test.

c. The P-value is the probability that the null hypothesis is true.

d. A statistically significant result means the P-value is “small.”

Exercises

E25. A random sample of dogs was checked to see how many wore a collar. The 95% confidence interval for the percentage of all dogs that wear a collar turned out to be from 0.82 to 0.96. Which of these is not a true statement?

A. You can reject the hypothesis that 75% of all dogs wear a collar.

B. You cannot reject the hypothesis that 90% of all dogs wear a collar.

C. If 90% of all dogs wear a collar, then you are reasonably likely to get a result like the one from this sample.

D. If 75% of all dogs wear a collar, then you are reasonably likely to get a result like the one from this sample.

E. The true proportion of dogs that wear a collar may or may not be between 0.82 and 0.96.
E26. A Gallup poll of a recent year estimated that 27% of adult Americans approved and 67% disapproved of the way Congress was handling its job. The margin of error was given as 3%. Which of these is not a true statement? [Source: Gallup, www.gallup.com, 2006.]

A. You can reject the hypothesis that 35% of adult Americans approved of the way Congress was handling its job on that date.
B. You cannot reject the hypothesis that 70% of all Americans disapproved of the way Congress was handling its job on that date.
C. The true proportion of all Americans who approved of the way Congress was handling its job on that date must lie between 0.24 and 0.30.
D. If, in fact, 25% of all Americans approved of the way Congress was handling its job on that date, then the observed sample result is reasonably likely.
E. If, in fact, 75% of all Americans disapproved of the way Congress was handling its job on that date, then the observed sample result is not reasonably likely.

E27. Suppose 22 students out of a random sample of 40 students carry a backpack to school. Follow steps a–c to test the claim that 60% of the students in the school carry backpacks to class.

a. State the null hypothesis in words and in symbols.
b. Calculate the value of the test statistic. Calculate the P-value for the test. Use this P-value in a sentence that explains what it represents.
c. What is your conclusion? Explain it in the context of this problem.

E28. Follow the three steps in E27 for a random sample of 40 students, of which 31 carry backpacks to class.

E29. The Gallup Organization asked 1003 adults to answer this question: “If you could pick only one specific problem for science to solve in the next 25 years, what would it be?” Thirty percent responded “a cure for cancer.” Use a test of significance to determine whether it is plausible that, if you could ask all adults in the United States this question, only 25% would pick a cure for cancer. [Source: Gallup, June 2001, www.publicagenda.org.]

E30. “Americans increasingly favor raising the driving age, a USA Today/CNN/Gallup poll has found. Nearly two-thirds—61%—say they think a 16-year-old is too young to have a driver’s license. Only 37% of those polled thought it was okay to license 16-year-olds, compared with 50% who thought so in 1995. A slight majority, 53%, thinks teens should be at least 18 to get a license. The poll of 1002 adults was conducted Dec. 17–19, 2004.” [Source: USA Today polls, Is 16 Too Young to Drive a Car? March 2, 2005.]

If you asked all adult residents of the United States if they thought that a 16-year-old is too young to have a driver’s license, is it plausible that only 55% would say yes?

E31. A Gallup poll asked 506 adult Americans this question: “Increased efforts by business and industry to reduce air pollution might lead to higher prices for the things consumers buy. Would you be willing to pay $500 more each year in higher prices so that industry could reduce air pollution, or not?” The report said, “Almost two-thirds (63%) claim they would be willing to pay $500 more per year for this purpose.” Following the four steps of a significance test, determine if there is convincing evidence that the true proportion of Americans who would be willing to pay $500 more for this purpose differs from two-thirds. [Source: Gallup, www.gallup.com, March 16, 2001.]

E32. In August 2005, a Harris poll asked, “Do you agree or disagree with this statement? Protecting the environment is so important that requirements and standards cannot be too high, and continuing environmental improvements must be made regardless of cost.” Of 1217 adults surveyed nationwide, 74% agreed, 24% disagreed, and 1%
were unsure. Following the four steps of a significance test, determine if this is convincing evidence that the true proportion of the nation’s adults who agree with the statement is different from two-thirds. Use \( \alpha = 0.01 \). [Source: The Harris Report, August 2005, www.pollingreport.com.]

E33. Suppose that when flipping a penny, the probability of getting heads is 0.5. Suppose 100 people perform a test of significance with this penny. Each person plans to reject the null hypothesis that \( p = 0.5 \) if their \( P \)-value is less than 0.05. How many of these people do you expect to make a Type I error?

E34. Fifty laboratories across the country are conducting similar tests on the possible side effects of a certain medication. Each laboratory tests the null hypothesis that no more than 15% of the subjects taking the medication develop undesirable side effects, and each test is done at the 10% significance level. One of these two questions can be answered. Answer that one, and explain why you cannot answer the other.
   a. How many of the laboratories do you expect to reject the null hypothesis if it is true?
   b. How many of the laboratories do you expect to reject the null hypothesis if it is false?

E35. Jeffrey and Taline each want to test whether the proportion of adults in their state who have graduated from college is 0.6, as claimed in the newspaper. Jeffrey takes a random sample of 200 adults and uses \( \alpha = 0.01 \). Taline takes a random sample of 200 adults and uses \( \alpha = 0.05 \). Suppose the newspaper’s percentage is actually right.
   a. Is it possible for Jeffrey or Taline to make a Type I error? If so, who is more likely to do so?
   b. Is it possible for Jeffrey or Taline to make a Type II error? If so, who is more likely to do so?

E36. Jeffrey and Taline each want to test whether the proportion of adults in their state who have gone to nursery school is 0.55, as claimed in the newspaper. Each takes a random sample of 200 adults. Jeffrey uses \( \alpha = 0.01 \). Taline uses \( \alpha = 0.05 \). Suppose the newspaper’s percentage is actually wrong.
   a. Is it possible for Jeffrey or Taline to make a Type I error? If so, who is more likely to do so?
   b. Is it possible for Jeffrey or Taline to make a Type II error? If so, who is more likely to do so?

E37. A friend wants to do a two-sided test of whether spinning pennies is fair. His null hypothesis is that the percentage of heads when a penny is spun is 50%. He plans to do 20 spins and use a 5% level of significance.
   a. Use the binomial probability formula on page 385 to construct the binomial distribution for \( n = 20 \) and \( p = 0.5 \).
   b. Use your distribution from part a to find the numbers of heads that would result in rejecting the null hypothesis.
   c. You are pretty sure that the true percentage of heads is about 40%. Supposing that is the case, use exact binomial probabilities to compute the probability that your friend will be able to reject the null hypothesis.
   d. What should you suggest to your friend?

E38. A quality improvement plan for a business office calls for selecting a sample of ten invoices that came in during a specified week and counting the number that have not been paid within 30 days. Initially, management wants to test to see if the proportion of invoices not paid on time exceeds 30%.
   a. Suppose the plan calls for rejecting the null hypothesis if the number of invoices not processed on time is six or greater out of the sampled ten. Use the binomial probability formula on page 385 to construct the binomial distribution for \( n = 10 \) and \( p = 0.3 \). Then use this distribution to find the significance level for this test.
   b. Suppose that, in fact, the true rate of invoices not processed on time is 40%. What is the power of the test described in
part a to detect this alternative? Use exact binomial probabilities to calculate this power.

E39. In P24 on page 510, you computed a test statistic for this study: Forty-five dogs and their owners, chosen at random, were photographed separately. A judge was shown a picture of each owner and pictures of two dogs and asked to pick the dog that went with the owner. The judge was right 23 times.

a. If you use a significance level of 0.05, do you reject the null hypothesis that, in the long run, the judge will select the correct dog half the time?

b. Suppose you compute the 95% confidence interval for the proportion of times the judge will select the correct dog. Will the value of $p_0$ from the null hypothesis be in this confidence interval? Explain.

c. Compute the confidence interval to check your answer to part b.

d. Explain how the margin of error for the confidence interval is related to the power of the test.

e. How would you suggest that the investigator get more power for this test?

E40. To perform a significance test for a proportion, you must decide on the hypotheses (null and alternative), the level of significance, and the sample size. Explain the effects of each of these components on the power of the test.

E41. At the beginning of this section, you read that 2% of barn swallows have white feathers in places where the plumage is normally blue or red. However, about 16% of barn swallows captured around Chernobyl after 1991 had such genetic mutations. The number captured around Chernobyl was relatively large—266 barn swallows.

a. If you want to determine whether the increase in the proportion of mutations is statistically significant, should this be a one-sided or two-sided test?

b. What condition(s) of the significance test for a proportion are not satisfied?

c. Display 8.19 shows the sampling distribution of the sample proportion, $\hat{p}$, for $n = 266$ and $p = 0.02$. Use it to estimate the $P$-value for the test you selected.

d. What should the researchers conclude?

E42. A psychologist was struck by the fact that at square tables many pairs of students sat on adjacent sides rather than across from each other and wondered if people prefer to sit that way. The psychologist collected some data, observing 50 pairs of students seated in the student cafeteria of a California university. The tables were square and had one seat available on each of the four sides. The psychologist observed that 35 pairs sat on adjacent sides of the table, while 15 pairs sat across from each other. He wanted to know if the evidence suggested a preference by students for adjacent sides, as opposed to simply random behavior. [Source: Joel E. Cohen, “Turning the Tables,” in Statistics by Example: Exploring Data, edited by Frederick Mosteller et al. (Reading, Mass.: Addison-Wesley, 1973), pp. 87–90.]
8.2 Testing a Proportion

a. What is the probability that if two students sit down randomly at a square table, they sit on adjacent sides?
b. What is an appropriate null hypothesis to test whether students’ seating pattern is random, as opposed to their having a preference to sit on adjacent sides?
c. What is the appropriate alternative hypothesis?
d. Does this situation meet the conditions for a significance test?
e. What is the test statistic?
f. Find the $P$-value for this test.
g. Write a conclusion for the psychologist.

E43. A psychology professor flashed a pair of head shots before subjects and asked them to identify the face that displayed more competence. The subjects didn’t know it, but the head shots were the two opposing candidates in 600 recent House of Representative elections. It turned out that 66.8% of the candidates who were perceived as more competent by the subjects were winners in the election. The article says that the statistical test used “tests the proportion of correctly predicted races against the chance level of 50%.” [Source: A. Todorov, A. N. Mandisodza, A. Goren, and C. C. Hall, “Inferences of Competence from Faces Predict Election Outcomes, Science 308 (2005): 1623–26.]

a. To determine the significance of this result, should a one-sided or two-sided test be used?
b. What is $\hat{p}$? What is $p_0$?
c. The test statistic given in the article is 8.25. Is this correct?
d. The $P$-value reported was “$P < 0.001$.” Is this correct?
e. Does the sample size make an important contribution to the power of the test in this situation? Explain.

E44. Suppose a test rejects this null hypothesis in favor of the alternative:

$H_0: p = p_0$ versus $H_1: p > p_0$

a. Explain why it also would have been reasonable to reject the null hypothesis if it had been of this form: $H_0: p < p_0$.
b. Would the null hypothesis necessarily have been rejected if the alternative had been $H_1: p \neq p_0$? Explain.

E45. Suppose that 42 out of 80 randomly selected students prefer hamburgers to hot dogs and that 38 prefer hot dogs to hamburgers.

a. Test the null hypothesis that the percentage of all students who prefer hamburgers is 0.55.
b. Test the null hypothesis that the percentage of all students who prefer hot dogs is 0.55.
c. How can you reconcile your conclusions in parts a and b?

Now suppose you go out and get a larger random sample of 800 students. Suppose that 420 of these students prefer hamburgers to hot dogs and that 380 prefer hot dogs to hamburgers.

d. Test the null hypothesis that the percentage of all students who prefer hamburgers is 0.55.
e. Test the null hypothesis that the percentage of all students who prefer hot dogs is 0.55.
f. How can you reconcile your conclusions in parts a, b, d, and e?
A Confi dence Interval for the Diff erence of Two Proportions

In Section 8.1, you found a confi dence interval estimate for the proportion of successes in a single population. The more common and important situation involves taking two samples independently from two different populations with the goal of estimating the size of the difference between the proportion of successes in one population and the proportion of successes in the other.

A recent poll of 29,700 U.S. households found that 63% owned a pet. The percentage in 1994 was 56%. [Source: American Pet Products Manufacturers Association, www.appma.org.] The two populations are the households in the United States in 1994 and the households now. The question you will investigate is “What was the change in the percentage of U.S. households that own a pet?” The obvious answer

<table>
<thead>
<tr>
<th>Year</th>
<th>Winning Candidate</th>
<th>Height</th>
<th>Runner-Up</th>
<th>Candidate</th>
<th>Height</th>
</tr>
</thead>
<tbody>
<tr>
<td>1900</td>
<td>McKinley</td>
<td>5'7&quot;</td>
<td>Bryan</td>
<td>6'0&quot;</td>
<td></td>
</tr>
<tr>
<td>1904</td>
<td>T. Roosevelt</td>
<td>5'10&quot;</td>
<td>Parker</td>
<td>6'0&quot;</td>
<td></td>
</tr>
<tr>
<td>1908</td>
<td>Taft</td>
<td>6'0&quot;</td>
<td>Bryan</td>
<td>6'0&quot;</td>
<td></td>
</tr>
<tr>
<td>1912</td>
<td>Wilson</td>
<td>5'11&quot;</td>
<td>T. Roosevelt</td>
<td>5'10&quot;</td>
<td></td>
</tr>
<tr>
<td>1916</td>
<td>Wilson</td>
<td>5'11&quot;</td>
<td>Hughes</td>
<td>5'11&quot;</td>
<td></td>
</tr>
<tr>
<td>1920</td>
<td>Harding</td>
<td>6'0&quot;</td>
<td>Cox</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>1924</td>
<td>Coolidge</td>
<td>5'10&quot;</td>
<td>Davis</td>
<td>6'0&quot;</td>
<td></td>
</tr>
<tr>
<td>1928</td>
<td>Hoover</td>
<td>5'11&quot;</td>
<td>Smith</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>1932</td>
<td>F. Roosevelt</td>
<td>6'2&quot;</td>
<td>Hoover</td>
<td>5'11&quot;</td>
<td></td>
</tr>
<tr>
<td>1936</td>
<td>F. Roosevelt</td>
<td>6'2&quot;</td>
<td>Landon</td>
<td>5'8&quot;</td>
<td></td>
</tr>
<tr>
<td>1940</td>
<td>F. Roosevelt</td>
<td>6'2&quot;</td>
<td>Wilkie</td>
<td>6'1&quot;</td>
<td></td>
</tr>
<tr>
<td>1944</td>
<td>F. Roosevelt</td>
<td>6'2&quot;</td>
<td>Dewey</td>
<td>5'8&quot;</td>
<td></td>
</tr>
<tr>
<td>1948</td>
<td>Truman</td>
<td>5'9&quot;</td>
<td>Dewey</td>
<td>5'8&quot;</td>
<td></td>
</tr>
<tr>
<td>1952</td>
<td>Eisenhower</td>
<td>5'10.5&quot;</td>
<td>Stevenson</td>
<td>5'10&quot;</td>
<td></td>
</tr>
<tr>
<td>1956</td>
<td>Eisenhower</td>
<td>5'10.5&quot;</td>
<td>Stevenson</td>
<td>5'10&quot;</td>
<td></td>
</tr>
<tr>
<td>1960</td>
<td>Kennedy</td>
<td>6'0&quot;</td>
<td>Nixon</td>
<td>5'11.5&quot;</td>
<td></td>
</tr>
<tr>
<td>1964</td>
<td>Johnson</td>
<td>6'3&quot;</td>
<td>Goldwater</td>
<td>6'0&quot;</td>
<td></td>
</tr>
<tr>
<td>1968</td>
<td>Nixon</td>
<td>5'11.5&quot;</td>
<td>Humphrey</td>
<td>5'11&quot;</td>
<td></td>
</tr>
<tr>
<td>1972</td>
<td>Nixon</td>
<td>5'11.5&quot;</td>
<td>McGovern</td>
<td>6'1&quot;</td>
<td></td>
</tr>
<tr>
<td>1976</td>
<td>Carter</td>
<td>5'9.5&quot;</td>
<td>Ford</td>
<td>6'0&quot;</td>
<td></td>
</tr>
<tr>
<td>1980</td>
<td>Reagan</td>
<td>6'1&quot;</td>
<td>Carter</td>
<td>5'9.5&quot;</td>
<td></td>
</tr>
<tr>
<td>1984</td>
<td>Reagan</td>
<td>6'1&quot;</td>
<td>Mondale</td>
<td>5'10&quot;</td>
<td></td>
</tr>
<tr>
<td>1988</td>
<td>G. Bush</td>
<td>6'2&quot;</td>
<td>Dukakis</td>
<td>5'8&quot;</td>
<td></td>
</tr>
<tr>
<td>1992</td>
<td>Clinton</td>
<td>6'2&quot;</td>
<td>G. Bush</td>
<td>6'2&quot;</td>
<td></td>
</tr>
<tr>
<td>1996</td>
<td>Clinton</td>
<td>6'2&quot;</td>
<td>Dole</td>
<td>6'2&quot;</td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>G. W. Bush</td>
<td>5'11&quot;</td>
<td>Gore</td>
<td>6'1&quot;</td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>G. W. Bush</td>
<td>5'11&quot;</td>
<td>Kerry</td>
<td>6'4&quot;</td>
<td></td>
</tr>
</tbody>
</table>

is that the percentage increased by 7 percentage points, but this is only an estimate because 7% is the difference of two sample percentages that probably are not exactly equal to the population percentages. In this section, you will learn how to find the margin of error to go with the difference, 7%.

A confidence interval for the difference of two proportions, $p_1 - p_2$, where $p_1$ is the proportion of successes in the first population and $p_2$ is the proportion of successes in the second population, has this familiar form:

\[
(\hat{p}_1 - \hat{p}_2) \pm z^* \cdot \text{standard error of } (\hat{p}_1 - \hat{p}_2)
\]

Here $\hat{p}_1$ and $\hat{p}_2$ are the proportions of successes in the two samples. Substituting what you have so far into the formula gives the 95% confidence interval for the difference between the proportion of U.S. households that own pets now and the proportion that owned pets in 1994:

\[
0.07 \pm 1.96 \cdot \text{standard error of } (\hat{p}_1 - \hat{p}_2)
\]

The hard part, as always, is estimating the size of the standard error.

**The Formula for the Confidence Interval**

The statistic of interest is a difference of random variables, and you know from Section 6.1 that the mean of a difference is the difference of the means. If the variables are independent, the variance of the difference is the sum of the variances. In symbols,

\[
\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2 \quad \text{and} \quad \sigma^2_{\hat{p}_1 - \hat{p}_2} = \sigma^2_{\hat{p}_1} + \sigma^2_{\hat{p}_2}
\]

In addition, under the independence assumption, the sampling distribution of the differences will be approximately normal if the sample sizes are suitably large.

Taking the square root of the variance, the standard error is

\[
\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\sigma^2_{\hat{p}_1} + \sigma^2_{\hat{p}_2}}
\]

From Section 7.3, the standard error of the distribution of the sample proportion $\hat{p}_1$ is

\[
\sigma_{\hat{p}_1} = \sqrt{\frac{p_1(1 - p_1)}{n_1}} \quad \text{which can be estimated by} \quad \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1}}
\]

where $n_1$ is the sample size. Similarly, the standard error of the distribution of the sample proportion $\hat{p}_2$ is

\[
\sigma_{\hat{p}_2} = \sqrt{\frac{p_2(1 - p_2)}{n_2}} \quad \text{which can be estimated by} \quad \sqrt{\frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}
\]

where $n_2$ is the sample size.
Substituting these into the formula for the standard error of a difference, you can estimate the standard error of the difference of two proportions:

$$\sigma_{\hat{p}_1 - \hat{p}_2} \approx \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

Now you can write the complete confidence interval for the difference between two proportions.

### Confidence Interval for the Difference of Two Proportions

The confidence interval for the difference, $p_1 - p_2$, of the proportion of successes in one population and the proportion of successes in a second population is

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \cdot \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

where $\hat{p}_1$ is the proportion of successes in a random sample of size $n_1$ taken from the first population and $\hat{p}_2$ is the proportion of successes in a random sample of size $n_2$ taken from the second population. (The sample sizes don’t have to be equal.)

The conditions that must be met in order to use this formula are that

- the two samples are taken randomly and independently from two populations
- each population is at least 10 times as large as its sample size
- $n_1\hat{p}_1$, $n_1(1 - \hat{p}_1)$, $n_2\hat{p}_2$, and $n_2(1 - \hat{p}_2)$ are all at least 5

### Example: A Difference in Pet Ownership?

As described at the beginning of this section, a recent pet ownership survey found that 63% of the 29,700 U.S. households sampled own a pet. A 1994 survey, taken by the same organization, found that 56% of the 6,786 U.S. households sampled owned a pet. Find and interpret a 95% confidence interval for the difference between the proportion of U.S. households that own a pet now and the proportion of U.S. households that owned a pet in 1994.

**Solution**

The two samples can be considered random samples. They were taken independently from the population of U.S. households in two different years. The number of U.S. households in each year is larger than 10 times 29,700. Finally, each of the following is at least 5.

$$n_1\hat{p}_1 = 29,700(0.63) = 18,711 \quad n_1(1 - \hat{p}_1) = 29,700(0.37) = 10,989$$

$$n_2\hat{p}_2 = 6,786(0.56) = 3,800 \quad n_2(1 - \hat{p}_2) = 6,786(0.44) \approx 2,986$$
The 95% confidence interval for the difference of the two population proportions, $p_1$ and $p_2$, is

$$
(\hat{p}_1 - \hat{p}_2) \pm z^* \cdot \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}
$$

$$
= (0.63 - 0.56) \pm 1.96 \sqrt{\frac{(0.63)(1 - 0.63)}{29,700} + \frac{(0.56)(1 - 0.56)}{6,786}}
$$

$$
= 0.07 \pm 0.013
$$

Alternatively, you can write this confidence interval as $(0.057, 0.083)$.

You also can calculate this interval using your calculator by entering each sample size and the number of successes in each sample. [See Calculator Note 8E.]

You are 95% confident that the difference in the two rates of pet ownership is between 0.057 and 0.083. This means that it is plausible that the difference in the percentage of households that own pets now and the percentage in 1994 is 5.7%. It is also plausible that the difference is 8.3%. Note that a difference of 0 does not lie within the confidence interval. This means that if the difference in the proportion of pet owners now and in 1994 actually is 0, getting a difference of 0.07 in the samples is not at all likely. Thus, you are convinced that there was a change in the percentage of households that own a pet.

It doesn’t matter which sample proportion you call $\hat{p}_1$ and which you call $\hat{p}_2$, as long as you remember which is which. Most people like to assign the larger one to be $\hat{p}_1$ so that the difference, $\hat{p}_1 - \hat{p}_2$, is positive. If you like working with negative numbers, do it the other way!

**DISCUSSION**

**The Formula for the Confidence Interval**

D41. A survey found that 45% of households own a dog, 34% own a cat, and 20% own both.

a. What percentage of households in the survey own neither a cat nor a dog?

b. Are owning a dog and owning a cat independent events?

c. Should you use the techniques of this section to estimate the difference in the percentage of households that own a dog and the percentage that own a cat? Explain.

Activity 8.3a will help you understand how confidence intervals for differences of proportions may vary from sample to sample.
Who's Yellow?

What you'll need: a sample of 50 Skittles, a sample of 50 Milk Chocolate M&M’s (or bags prepared by your instructor)

1. Count the number of yellow candies in your sample of Skittles and in your sample of M&M’s. Compute the two sample proportions.
2. Find the 95% confidence interval for the difference in the two sample proportions:

   \[ \frac{\text{proportion of Skittles that are yellow}}{\text{proportion of M&M’s that are yellow}} \]

   (Every student should subtract in this order, even if the difference is negative.)

3. Draw your confidence interval on a large copy of Display 8.21. Other members of your class should draw their confidence intervals on the same chart.

4. Compare the centers of the confidence intervals. Are they close to each other, or do they vary a lot? Why do they vary?

5. Compare the widths of the confidence intervals. Are the widths fairly equal, or do they vary a lot? Why is there any variability in the width?

6. How many of the confidence intervals would you expect to capture the true difference in proportions?

7. Skittles have 20% yellow candies. Milk Chocolate M&M’s have 14% yellow candies. How many of the confidence intervals actually captured the true difference in these proportions?
Summary 8.3: A Confidence Interval for the Difference of Two Proportions

When the sample sizes, \( n_1 \) and \( n_2 \), are large enough, the sampling distribution of \( \hat{p}_1 - \hat{p}_2 \) is approximately normal, is centered at \( p_1 - p_2 \), and has standard error

\[
\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}
\]

In practice, you estimate \( \sigma_{\hat{p}_1 - \hat{p}_2} \) using the two sample proportions. This leads to a confidence interval for the difference of the two population proportions of

\[
(\hat{p}_1 - \hat{p}_2) \pm z^* \cdot \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}
\]

where \( \pm z^* \) are the values that enclose an area in a normal distribution equal to the confidence level, which typically is 0.95. Don’t forget to check the three conditions on page 520 that must be met to use this confidence interval, especially the condition that the two samples were taken independently from two different populations.

When interpreting, say, a 95% confidence interval, you should say that you are 95% confident that the difference between the proportion of successes in the first population and the proportion of successes in the second population is in this confidence interval. (However, say this in the context of the situation.) If the confidence interval includes 0, you can’t conclude that there is any difference in the proportions in the two populations.

Practice

The Formula for the Confidence Interval

P46. Suppose the surveys on pet ownership described in the example on page 520 had used sample sizes of only 100 people. That is, a 1994 survey of 100 U.S. households found that 56% owned a pet, and this year a survey of 100 U.S. households found that 63% owned a pet.

a. Check the conditions for constructing a confidence interval for the difference of two proportions.

b. Compute and interpret the confidence interval.

c. Is 0 in the confidence interval? What does your answer imply?

P47. A poll of 549 teenagers asked if it was appropriate for parents to install a special device on the car to allow parents to monitor teenagers’ driving speeds. Among the 325 13–15-year-olds in the sample, 50% responded yes to this question. Among the 224 16–17-year-olds, only 28% responded yes. [Source: Gallup, Teens Slow to Support Speed Monitors, 2005, www.poll.gallup.com.]

a. Check the conditions for constructing a confidence interval for the difference of two proportions.

b. Estimate the difference between the proportion of all 13–15-year-olds who would answer yes to this question and the proportion of all 16–17-year-olds who would answer yes.

c. Interpret the resulting interval in the context of the problem.

d. Is 0 in the confidence interval? What does your answer imply?
Exercises

E47. A statistics instructor kept track of students’ grades on her final examination and whether they did the exam in pencil or in pen. Of the 75 students who used a pencil, 63 got grades of 70% or higher. Of the 19 students who used a pen, 10 got grades of 70% or higher. Her 95% confidence interval for the difference is (0.07, 0.55). Assume that these students can be considered random samples from the students who could take this instructor’s final examination. Which of these choices are the best two interpretations of this confidence interval?
A. The teacher must observe more students, because this confidence interval is too wide to be of any possible use.
B. The teacher has evidence that a larger proportion of students who use a pencil score 70% or higher on her final than students who use a pen.
C. If the teacher could observe all students who could take her final examination, she is 95% confident that the difference in the proportion who use a pencil and get a grade of 70% or higher and the proportion who use a pen and get a grade of 70% or higher is between 0.07 and 0.55.
D. If the teacher could observe all students who could take her final examination, she is 95% confident that between 7% and 55% of them will get 70% or higher.

E48. According to the American Automobile Association, a study found that 85.8% of 207,638 drivers who were involved in a fatal car accident had a valid driver’s license. Of 16,226 motorcycle drivers involved in a fatal accident, 80.1% had a valid driver’s license. A 95% confidence interval for the difference in proportions is (0.05, 0.06). You can consider these random samples of drivers involved in fatal accidents. Which of these choices are the best two interpretations of this confidence interval? [Source: AAA Foundation for Traffic Safety, Unlicensed to Kill: The Sequel, 2003, www.aaafoundation.org.]
A. You can be confident that, if you looked at all drivers involved in fatal accidents, a larger percentage of drivers of cars would have a valid license than drivers of motorcycles.
B. If all drivers involved in a fatal accident were observed, you can be 95% confident that the true percentage of drivers who don’t have a driver’s license would be between 5% and 6%.
C. This confidence interval is so narrow because the sample sizes are so big.
D. If all drivers who don’t have a valid driver’s license were observed, you are 95% confident that between 5% and 6% more of the unlicensed drivers of motorcycles would have a fatal accident than unlicensed drivers of cars.

E49. Researchers at the University of California, San Diego, found that when given a choice of two dogs, judges correctly matched 16 out of 25 purebred dogs with their owners. When the dogs were mutts, only 7 dogs out of 20 were correctly matched with their owners. [Source: www.newscientist.com.]

a. Check the conditions for constructing a confidence interval for the difference of two proportions.
b. Find a 95% confidence interval for the difference of proportions.
c. Interpret the resulting interval in the context of the problem.
d. Is 0 in the confidence interval? What does your answer imply?
e. What would you recommend that the researchers do?
E50. The USC Annenberg School Center for the Digital Future found that 66.9% of Americans used the Internet in 2000 and 78.6% used the Internet in 2005. Assume that the samples were independently and randomly selected and that the sample size was 2000 in both years. [Source: www.digitalcenter.org.]

a. Check the conditions for constructing a confidence interval for the difference of two proportions.

b. Find a 99% confidence interval for the difference of proportions.

c. Interpret the resulting interval in the context of the problem.

d. Is 0 in the confidence interval? What does your answer imply?

E51. A Harris poll asked this question of about 425 men and 425 women: “When you get a sales or customer service phone call from someone you don’t know, would you prefer to be addressed by your first name or by your last name, or don’t you care one way or the other?” Twenty-three percent of the men and 34% of the women said they would prefer being addressed by their last name. Although this was not quite the case, you can assume that the samples were random and independent. [Source: Harris, November 14, 2001, www.harrisinteractive.com.]

a. Check the conditions for constructing a confidence interval for the difference of two proportions.

b. Find a 99% confidence interval for the difference of proportions.

c. Interpret the resulting interval in the context of the problem.

d. Is 0 in the confidence interval? What does your answer imply?

E52. “At what age do you think people should be permitted to have a driver’s license?” In 1995, 46% of 1000 randomly sampled adults in the United States responded that 16 was the correct age. In 2004, only 35% of 1000 randomly sampled adults responded this way. [Source: USA Today/CNN/Gallup poll, December 17–19, 2004.]

a. Check the conditions for constructing a confidence interval for the difference of two proportions.

b. Estimate the true difference between the 1995 and 2004 proportions favoring 16 as the correct age to begin driving.

c. Interpret the resulting interval in the context of the problem.

d. Is 0 in the confidence interval? What does your answer imply?

E53. In a recent national survey, 15,200 high school seniors in 128 schools completed questionnaires about physical activity. Male students were significantly more likely (59%) than female students (48%) to have played on sports teams run by their school during the 12 months preceding the survey. Check the accuracy of the phrase “significantly more likely,” assuming that there were equal numbers of male and female students in this survey and that the samples are equivalent to independent simple random samples. [Source: Participation in School Athletics, www.childtrendsdatabank.org.]

8.3 A Confidence Interval for the Difference of Two Proportions 525
E54. In the report referenced in E53, the participation of females in school athletics increased from 47% in 1991 to 48% recently. The total sample size (males and females) for 1991 was approximately 15,400. Does this increase represent a significant change in the level of participation for females? Explain your reasoning.

E55. What is the effect of an increase in the sample sizes on the width of a confidence interval for the difference of two proportions?

E56. Statements I and II are interpretations of a 95% confidence interval for a single proportion. Write similar statements for a 95% confidence interval for a difference of two proportions.
   I. A 95% confidence interval consists of those population proportions \( p \) for which the proportion from the sample, \( \hat{p} \), is reasonably likely to occur.
   II. If you construct a hundred 95% confidence intervals, you expect that the population proportion \( p \) will be in 95 of them.

E57. The formula for the confidence interval for the difference of two proportions involves \( z \). Why is it okay to use \( z \) in this case?

E58. What is suggested if the confidence interval for the difference of two proportions
   a. includes 0?
   b. does not include 0?

E59. A Harris online poll asked 99 online respondents, "Would you support the proposed bill that would eliminate the U.S. penny?" There were 22 (or 22.2%) yes votes and 77 (or 77.8%) no votes. [Source: February 25, 2002, www.harriszone.com.] A student computed a 95% confidence interval for the difference of the proportion of the population answering yes and the proportion answering no as follows:

\[
(\hat{p}_1 - \hat{p}_2) \pm z^* \cdot \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \\
= (0.222 - 0.778) \\
\pm 1.96 \sqrt{\frac{(0.222)(0.778)}{99} + \frac{(0.778)(0.222)}{99}} \\
= -0.556 \pm 0.116
\]

Assuming that the respondents can be considered a random sample from some population, is the student's method correct? If so, write an interpretation of this interval. If not, do a more appropriate analysis.

E60. In the USA Today poll on driving ages described in E52, 35% of 1000 randomly sampled adults chose 16 as the preferred initial driving age while 42% chose 18. Can the difference of the proportion in the population who prefer 16 and the proportion in the population who prefer 18 be estimated by the confidence interval developed in this chapter? Explain why or why not.

---

**A Significance Test for the Difference of Two Proportions**

In the previous section, you learned how to estimate the size of the difference of two proportions. But sometimes you must decide between two alternatives. For example:

- Are snowboarders or skiers more likely to be seriously injured?
- Have opinions about the general state of the country changed over the past 5 years?
- Is there a difference between the proportions of teen boys and girls who favor dress codes?
The size of any difference involved is not the issue, as it was in Section 8.3. All you care about is whether you have enough evidence to conclude that there is a difference. In this section, you will learn to perform a test of significance in order to decide if the observed difference can reasonably be attributed to chance alone or if the difference is so large that something other than chance variation must be causing it.

**A Sampling Distribution of the Difference**

Before you learn how to use a test of significance to answer questions like those on the previous page, you will learn about what sizes of differences you should expect in your samples when there is no actual difference in the proportions of successes in the two populations. For example, as candy connoisseurs are well aware, there is no difference in the percentage of orange candies in Skittles and in Milk Chocolate M&M’s. Yet, when you take random samples of each kind of candy, you usually get a difference in the two proportions of orange candies in your samples. In Activity 8.4a, you will examine what a sampling distribution of \( \hat{p}_1 - \hat{p}_2 \) looks like when \( p_1 = p_2 \).

**ACTIVITY 8.4a**

**Differences in Proportions of Orange**

**What you’ll need:** a sample of 50 Skittles, a sample of 50 Milk Chocolate M&M’s (or bags prepared by your instructor)

1. Count the number of orange candies in your sample of Skittles and in your sample of M&M’s. Compute the two sample proportions.
2. Compute the difference in the two sample proportions:

\[
\frac{\text{proportion of Skittles that are orange}}{\text{proportion of M&M’s that are orange}}
\]

(Every student should subtract in this order, even if the difference is negative.)

3. Each member of your class should mark his or her difference in sample proportions above a number line so that everyone can see the distribution.
4. Repeat with fresh samples until you have at least 100 differences.
5. What is the shape of the sampling distribution?
6. Where should the sampling distribution be centered? Where is it centered?
7. What, approximately, is the standard error of the sampling distribution?
Approximate sampling distributions (constructed from 5000 pairs of samples) for samples of size 30, 50, and 100 are shown in Display 8.22. In each case, the proportion of successes in the first population is 0.2, and the proportion of successes in the second population is 0.2.

Display 8.22  Simulated sampling distributions of \( \hat{p}_1 - \hat{p}_2 \) when \( p_1 = p_2 = 0.2 \) for various sample sizes.

Note three facts about these approximate sampling distributions:
- Each sampling distribution is approximately normal in shape and becomes more so with larger sample sizes.
- The mean of each sampling distribution is at \( p_1 - p_2 = 0.2 - 0.2 = 0 \).
- The formula from Section 8.3 can be used to find the standard error:

\[
\sigma_{\hat{p}_1 - \hat{p}_2} \approx \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}} \approx \sqrt{\frac{0.2(1 - 0.2)}{n_1} + \frac{0.2(1 - 0.2)}{n_2}}
\]

Display 8.23 lists the SEs computed from the formula and from the approximate sampling distributions in Display 8.22. They match very closely.
A Significance Test for the Difference of Two Proportions

Sample Size for \( n_1 \) and \( n_2 \)

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>( SE ) when ( p_1 = p_2 = 0.2 )</th>
<th>( SE ) from Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>0.103</td>
<td>0.104</td>
</tr>
<tr>
<td>50</td>
<td>0.080</td>
<td>0.079</td>
</tr>
<tr>
<td>100</td>
<td>0.057</td>
<td>0.058</td>
</tr>
</tbody>
</table>

Display 8.23  \( SEs \) from the formula and from the simulation.

A Sampling Distribution of the Difference

D42. Use your calculator or computer to take 200 pairs of random samples, both samples of size 25 and both from populations with 35% successes.

a. Construct a plot of the differences of the two sample proportions.
b. From your plot, estimate the mean and standard error of the sampling distribution of \( \hat{p}_1 - \hat{p}_2 \).
c. Find the theoretical value of the mean and standard error of the sampling distribution of \( \hat{p}_1 - \hat{p}_2 \). Compare your answers to your estimates in part b.

The Theory of a Significance Test for the Difference of Two Proportions

Let’s see how these ideas apply to a real sample survey. Think about the left-handers you know or know about. Queen Victoria was left-handed, and so were Julius Caesar, Leonardo da Vinci, and Benjamin Franklin. Angelina Jolie and Oprah Winfrey are left-handed, and so are Bruce Willis, Keanu Reeves, and Spike Lee. You probably know more left-handed males than females. So you might ask this question: “Is there objective evidence to show that the proportion of left-handed males in the country is larger than the proportion of left-handed females?” In fact, there are objective data that can help us answer this question.

The government’s Health and Nutrition Survey (HANES) of 1976–80 recorded the handedness of random samples of 1067 males and 1170 females from across the country. The study found that 113 of the males and 92 of the females were left-handed. From these data you can construct a significance test to answer the question posed above. [Source: Freedman, Pisani, and Purves, Statistics, 3rd ed. (Norton, 1998), p. 537.]

First, a few words about the HANES survey are in order. The survey did not randomly sample 1067 males and then randomly sample 1170 females to ask about handedness. It randomly (by a somewhat complex design) selected a large sample of adults from across the country and then sorted out the males and females. The resulting two samples still can be considered to be independent and random, as long as the sample sizes are large.

Example: Are Men More Likely to Be Left-Handed?

The survey just described found that 113 of 1067 males were left-handed while 92 of 1170 females were left-handed. Perform a significance test to determine whether males are more likely to be left-handed than females. (Conditions will be checked in D43.)
Solution

Let \( p_1 \) denote the proportion of left-handers among all males in the adult population, and let \( p_2 \) denote the proportion among all adult females. The null hypothesis is almost always one of "no difference" or "no effect," so the appropriate null hypothesis here is

\[ H_0: \text{There is no difference between the proportions of left-handed males and left-handed females. In symbols, } p_1 = p_2, \text{ or } p_1 - p_2 = 0. \]

The alternative hypothesis, sometimes called the research hypothesis, is a statement of what the researcher is trying to establish. Here you are looking for evidence that the proportion of left-handers is greater among males, so the alternative hypothesis is

\[ H_a: \text{The proportion of left-handers among male adults is greater than the proportion among female adults. In symbols, } p_1 > p_2, \text{ or } p_1 - p_2 > 0. \]

The test statistic builds on the best estimates of these population proportions, namely, the sample proportions

\[ \hat{p}_1 = \frac{113}{1067} \approx 0.106 \text{ and } \hat{p}_2 = \frac{92}{1170} = 0.079 \]

The general form of a test statistic for testing hypotheses is

\[
\frac{\text{statistic} - \text{parameter}}{\text{standard deviation of statistic}}
\]

The difference from the sample (statistic) is \( \hat{p}_1 - \hat{p}_2 \), or \( 0.106 - 0.079 = 0.027 \). The hypothesized difference (parameter) is 0, the value under the null hypothesis \( p_1 - p_2 = 0 \). The standard error of the estimate is given exactly by

\[
\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}
\]

You could do as you did in Section 8.3: Estimate \( p_1 \) with \( \hat{p}_1 \) and estimate \( p_2 \) with \( \hat{p}_2 \). However, you can do even better. The null hypothesis states that the proportion of males who are left-handed is equal to the proportion of females who are left-handed, that is, \( p_1 = p_2 \). You can estimate this common value of \( p_1 \) and \( p_2 \) by combining the data from both samples into a pooled estimate, \( \hat{p} \). This pooled estimate is found by combining males and females into one group:

\[
\hat{p} = \frac{\text{total number of left-handers}}{\text{total number of people}} = \frac{113 + 92}{1067 + 1170} = \frac{205}{2237}
\]

The standard error of the difference then is approximately

\[
\sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}} \approx \sqrt{\frac{\hat{p}(1 - \hat{p})}{n_1} + \frac{\hat{p}(1 - \hat{p})}{n_2}}
\]

\[
= \sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}
\]
The test statistic then takes on the value

\[ z = \frac{\text{statistic} - \text{parameter}}{\text{standard deviation of statistic}} \]

\[ = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1 - \hat{p}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \]

\[ = \frac{\left( \frac{113}{1067} - \frac{92}{1170} \right) - 0}{\sqrt{\frac{205}{2237} \left( 1 - \frac{205}{2237} \right) \left( \frac{1}{1067} + \frac{1}{1170} \right)}} \approx 2.23 \]

This test statistic is based on the difference of two approximately normally distributed random variables, \( \hat{p}_1 \) and \( \hat{p}_2 \). You learned in Chapter 7 that such a difference itself has a normal distribution. Thus, you can use Table A on page 824 to find the P-value. The test statistic \( z = 2.23 \) has a (one-sided) P-value of 0.0129.

Display 8.24  P-value for a one-sided test with \( z = 2.23 \).

This P-value is very small, so reject the null hypothesis. The difference between the rates of left-handedness in these two samples is too large to attribute to chance variation alone. The evidence supports the alternative that males have a higher rate of left-handedness.

As you have seen, the significance test for the difference of two proportions proceeds along the same lines as the test for a single proportion. The steps in the significance test for the difference of two proportions are given in this box.

Components of a Significance Test for the Difference of Two Proportions for Surveys

1. **Check conditions.**

   A random sample of size \( n_1 \) is taken from a large binomial population with proportion of successes \( p_1 \). The proportion of successes in this sample is \( \hat{p}_1 \). A second and independent random sample of size \( n_2 \) is taken from a large binomial population with proportion of successes \( p_2 \). The proportion of successes in this sample is \( \hat{p}_2 \). All these quantities must be at least 5:

\[
 n_1\hat{p}_1 \quad n_1(1 - \hat{p}_1) \quad n_2\hat{p}_2 \quad n_2(1 - \hat{p}_2) \]

Each population size must be at least 10 times the sample size.
2. Write a null and an alternative hypothesis.
   You can write the null hypothesis in several different ways as long as you define $p_1$ and $p_2$:
   \[ H_0: \text{The proportion of successes } p_1 \text{ in the first population is equal to the proportion of successes } p_2 \text{ in the second population.} \]
   \[ H_0: p_1 = p_2, \text{ where } p_1 \text{ is the proportion of successes in the first population and } p_2 \text{ is the proportion of successes in the second population.} \]
   \[ H_0: p_1 - p_2 = 0, \text{ where } p_1 \text{ is the proportion of successes in the first population and } p_2 \text{ is the proportion of successes in the second population.} \]

   The form of the alternative hypothesis depends on whether you have a two-sided or a one-sided test:
   \[ H_a: \text{The proportion of successes } p_1 \text{ in the first population is not equal to the proportion of successes } p_2 \text{ in the second population. In symbols, } p_1 \neq p_2 \text{ or } p_1 - p_2 \neq 0. \]
   \[ H_a: \text{The proportion of successes } p_1 \text{ in the first population is greater than the proportion of successes } p_2 \text{ in the second population. In symbols, } p_1 > p_2 \text{ or } p_1 - p_2 > 0. \]
   \[ H_a: \text{The proportion of successes } p_1 \text{ in the first population is less than the proportion of successes } p_2 \text{ in the second population. In symbols, } p_1 < p_2 \text{ or } p_1 - p_2 < 0. \]

3. Compute the test statistic and a $P$-value.
   \[ z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \]
   where $\hat{p} = \frac{\text{total number of successes in both samples}}{n_1 + n_2}$

   The $P$-value is the probability of getting a value of $z$ as extreme as or even more extreme than that from your samples if $H_0$ is true. See Display 8.25.

4. Write a conclusion.
   State whether you reject or do not reject the null hypothesis. Link this conclusion to your computations by comparing $z$ to the critical value $z^*$ or by appealing to the $P$-value. You reject the null hypothesis if $z$ is more extreme than $z^*$ or, equivalently, if the $P$-value is smaller than the level of significance, $\alpha$. Then write a sentence giving your conclusion in context.
Example: Significance Test for a Difference

Should schools be allowed to restrict what students wear? A random sample of 548 teens, about equally split between boys and girls, was asked this question. A report on the results of the poll stated: “Gender seems to play a role in teens’ opinions about dress codes. While a slight majority of girls (57%) support restrictions on school wear, fewer boys (42%) agree.” [Source: Gallup, Youth Survey: Should Schools Police Fashion?, 2005, poll.gallup.com.] Is this evidence of a difference between the responses of boys and girls in the population sampled in this poll? Perform a test of significance at the 1% significance level.

Solution

1. Although Gallup uses a sampling method more complex than simple random sampling, you can consider these to be independent random samples. (As mentioned previously, dividing one random sample into two groups—boys and girls—gives an approximately random sample from each group when the sample sizes are large.) The proportion of girls in the sample who support restrictions is \( \hat{p}_1 = 0.57 \), and the proportion of boys in the sample who support restrictions is \( \hat{p}_2 = 0.42 \). Using the estimates of \( n_1 = 274 \) and \( n_2 = 274 \), you can see that all of

\[
\text{n}_1 \hat{p}_1 = 274(0.57) \approx 156 \quad \text{n}_1(1 - \hat{p}_1) = 274(1 - 0.57) \approx 118
\]
\[
\text{n}_2 \hat{p}_2 = 274(0.42) \approx 115 \quad \text{n}_2(1 - \hat{p}_2) = 274(1 - 0.42) \approx 159
\]

are larger than 5. Both the number of teenage girls and the number of teenage boys in the United States are much larger than 10 times each sample size.

2. \( H_0: \) The proportion \( p_1 \) of teenage girls supporting dress codes is equal to the proportion \( p_2 \) of teenage boys supporting dress codes.

\( H_a: p_1 \neq p_2. \) (Note that you can use the symbols \( p_1 \) and \( p_2 \) because you defined exactly what they stand for in the null hypothesis.)

3. The pooled estimate is

\[
\hat{p} = \frac{\text{total number of successes in both samples}}{n_1 + n_2} = \frac{156 + 115}{274 + 274} = 0.495
\]

The test statistic is

\[
z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \approx \frac{(0.57 - 0.42) - (0)}{\sqrt{0.495(1 - 0.495)\left(\frac{1}{274} + \frac{1}{274}\right)}} \approx 3.51
\]

With a test statistic of 3.51, the \( P \)-value for a two-sided test is 0.0002(2), or 0.0004. See Display 8.26 on the next page.
Chapter 8  Inference for Proportions

The proportion of girls supporting dress codes was 0.57 (56 out of 98 respondents), and the proportion of boys supporting dress codes was 0.41 (46 out of 112 respondents). This difference is statistically significant, and you reject the null hypothesis. If the proportions of girls and boys supporting dress codes were equal, then you would expect only 4 out of 10,000 repeated samples of size 274 to have a difference in sample proportions of 15% or larger. Because the P-value, 0.0004, is less than \( \alpha = 0.01 \), you cannot reasonably attribute the difference to chance variation. You conclude that the proportions of girls and boys supporting dress codes are different.

Display 8.26  A two-sided test with \( z = 3.51 \).

Note that you can use your calculator’s two-proportion \( z \)-test, which carries more decimal places, to get the more accurate value \( z = 3.503 \) and a \( P \)-value of 0.00046. [See Calculator Note 8F.]

Conclusion in context

4.  This difference is statistically significant, and you reject the null hypothesis. If the proportions of girls and boys supporting dress codes were equal, then you would expect only 4 out of 10,000 repeated samples of size 274 to have a difference in sample proportions of 15% or larger. Because the \( P \)-value, 0.0004, is less than \( \alpha = 0.01 \), you cannot reasonably attribute the difference to chance variation. You conclude that the proportions of girls and boys supporting dress codes are different.

Display 8.27  shows a printout for this example.

Test of Girls v Boys

Attribute (categorical): Girls
Attribute (categorical or grouping): Boys

In Girls 156 out of 274, or 0.569343, are 1
In Boys 115 out of 274, or 0.419708, are 1

Alternative hypothesis: The population proportion for 1 in Girls is not equal to that for 1 in Boys

The test statistic, \( z \), is 3.503.

If it were true that the two population proportions were equal (the null hypothesis), and the sampling process were performed repeatedly, the probability of getting a value of \( z \) with an absolute value this great or greater would be 0.00046.

Note: This probability was computed using the normal approximation.

Display 8.27  A printout for the significance test, with supporting response coded 1.
### The Theory of a Significance Test for the Difference of Two Proportions

D43. In the left-handedness example on page 529, the null hypothesis was rejected.

a. Which type of error might have been made in the significance test? Is the chance of making this error very large?

b. How can you check quickly that the condition is met that all of \( n_1 \hat{p}_1, n_2(1 - \hat{p}_1), n_1\hat{p}_2, \) and \( n_2(1 - \hat{p}_2) \) are at least 5?

D44. What does it suggest about the two proportions if the significance test for a difference of two proportions

a. fails to reject the null hypothesis?  
b. rejects the null hypothesis?

D45. Suppose one population has 10% successes and a second population has 20% successes. Generate an approximate sampling distribution of the difference between sample proportions for samples of size 10 from each population. Now repeat the procedure for samples of size 50 from each population. Explain how the two sampling distributions illustrate the need for the condition that \( n_1 \hat{p}_1, n_1(1 - \hat{p}_1), n_2\hat{p}_2, \) and \( n_2(1 - \hat{p}_2) \) must each be at least 5.

### Summary 8.4: A Significance Test for the Difference of Two Proportions

To get a simulated sampling distribution of the difference \( \hat{p}_1 - \hat{p}_2 \), you

- take independent random samples from two different populations with proportions of success \( p_1 \) and \( p_2 \)
- compute the proportion of successes in each sample, \( \hat{p}_1 \) and \( \hat{p}_2 \)
- subtract the proportions of successes in the samples, that is, find \( \hat{p}_1 - \hat{p}_2 \)
- repeat this process many times

You learned how to test whether two samples were drawn from populations that have the same proportion of successes. Follow these steps as with any significance test:

- **Check the conditions.** You need two sufficiently large random samples selected independently from two different populations.
- **Write a null and an alternative hypothesis.** The null hypothesis typically is that the two populations contain the same proportion of “successes.”
- **Compute the test statistic and a P-value.** The test statistic is

\[
    z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1 - \hat{p})\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}
\]

where \( \hat{p} \) is computed by combining both samples.

The P-value is the probability of getting a value of \( z \) as extreme as or even more extreme than that from your samples if \( H_0 \) is true.

- **Write a conclusion.** State whether you reject or do not reject the null hypothesis, linking this decision to your value of \( z \) or to your P-value. Then restate your conclusion in the context of the situation.
A Sampling Distribution of the Difference

P48. According to the 2006 *Statistical Abstract of the United States*, 12% of American adults visited the zoo in the last year. Suppose you take a random sample of 1000 American adults and compute the proportion, \( \hat{p}_1 \), who went to the zoo last year. Then you take another random sample of 800 American adults and compute the proportion, \( \hat{p}_2 \), in that sample who went to the zoo last year. You subtract the two proportions to get \( \hat{p}_1 - \hat{p}_2 \). You repeat this process until you have a distribution of millions of values of \( \hat{p}_1 - \hat{p}_2 \). (That’s why the problem says “suppose.”)

a. What is the expected value of the difference \( \hat{p}_1 - \hat{p}_2 \)?
b. What is the standard error of the distribution of \( \hat{p}_1 - \hat{p}_2 \)?
c. Sketch the distribution of \( \hat{p}_1 - \hat{p}_2 \), with a scale on the horizontal axis.
d. Compute the probability that the difference will be 0.05 or greater.

d. Theoretically, the plot should be centered exactly at 0.
e. To have a greater chance of having the difference \( \hat{p}_1 - \hat{p}_2 \) nearer 0, have larger sample sizes.

The Theory of a Significance Test for the Difference of Two Proportions

P50. A random sample of 600 probable voters was taken 3 weeks before the start of a campaign for mayor, and 321 of the 600 said they favored the new candidate over the incumbent. However, it was revealed the week before the election that the new candidate had dozens of outstanding parking tickets. Subsequently, a new random sample of 750 probable voters showed that 382 voters favored the new candidate. Do these data support the conclusion that there was a decrease in voter support for the new candidate after the parking tickets were revealed? Give appropriate statistical evidence to support your answer.

P51. What type of error might have been made in the significance test in the example on page 533 about teens supporting a dress code?
this process until you have a distribution of millions of values of \( \hat{p}_1 - \hat{p}_2 \).

a. What is the expected value of \( \hat{p}_1 - \hat{p}_2 \)?
b. What is the standard error of the distribution of \( \hat{p}_1 - \hat{p}_2 \)?
c. Sketch the distribution of \( \hat{p}_1 - \hat{p}_2 \), with a scale on the horizontal axis.
d. Compute the probability that the difference will be 0.05 or greater.

E63. A study of newly diagnosed lung cancer patients who had enrolled in an Early Lung Cancer Action Program regimen of screening found that 234 of the 258 patients with tumors measuring 15 mm or less had no metastases (spread to other parts of the body) and that 98 of the 118 patients with tumors measuring 16–25 mm had no metastases. The researchers needed to determine whether the difference in these two proportions is statistically significant. [Source: “Computed Tomographic Screening for Lung Cancer: The Relationship of Disease Stage to Tumor Size,” Archives of Internal Medicine 166 (February 13, 2006): 321–25.]

a. The investigators chose to do a one-sided test of significance. Why does that make sense?
b. Are the conditions for inference met?
c. The \( P \)-value reported in the article was 0.02. Is that correct?
d. Interpret this \( P \)-value in the context of this situation.

e. Based on using a one-sided test, is this a statistically significant difference? That is, if all teens were asked, are you confident that a larger proportion of girls than boys would answer yes? Assume that the samples were selected randomly.

b. The report says, “Participants were selected through random digit dialing.” Do you have any concerns about whether such a procedure would give a random sample?

c. Find a 95% confidence interval for the proportion of all teens who would answer yes. What additional assumption do you need to make to do this?

E64. A study at UCLA found that 55% of 100 cosmetic surgery patients take herbal supplements while only 24% of a similar group of 100 people not undergoing surgery take herbal supplements. The researchers needed to determine whether the difference in these two proportions is statistically significant. [Source: Los Angeles Times, February 20, 2006, page F4.]

a. Would you have chosen to do a one-sided or a two-sided test of significance? Explain your reasons.
b. Are the conditions for inference met?
c. Using the test you chose in part a, find the \( P \)-value for a significance test of the difference in the two proportions.
d. Interpret the \( P \)-value you found in part c in the context of this situation.

E65. A poll of 256 boys and 257 girls ages 12–17 asked, “Do you feel like you are personally making a positive difference in your community?” More girls (76%) than boys (63%) answered yes. [Source: www.ncpc.org, April 17, 2002.]

a. What is the expected value of \( \hat{p}_1 - \hat{p}_2 \)?
b. What is the standard error of the distribution of \( \hat{p}_1 - \hat{p}_2 \)?
c. Sketch the distribution of \( \hat{p}_1 - \hat{p}_2 \), with a scale on the horizontal axis.
d. Compute the probability that the difference will be 0.05 or greater.

e. Based on using a one-sided test, is this a statistically significant difference? That is, if all teens were asked, are you confident that a larger proportion of girls than boys would answer yes? Assume that the samples were selected randomly.

b. The report says, “Participants were selected through random digit dialing.” Do you have any concerns about whether such a procedure would give a random sample?

c. Find a 95% confidence interval for the proportion of all teens who would answer yes. What additional assumption do you need to make to do this?

E66. An annual poll of the sleeping habits of Americans found that of a random sample of 177 Americans ages 18–29, 30% slept 8 hours or more on a weekday. Of a random sample of 616 people ages 30–49, 24% slept 8 hours or more on a weekday. Is this a statistically significant difference? (When answering this question, include all four steps in your test of significance.) [Source: National Sleep Foundation, 2005, www.sleepfoundation.org.]

E67. In 2005, as NASA prepared to launch the New Horizons probe that will fly to Pluto on a 9-year mission, a poll asked 1000 adult Americans about their attitude toward the U.S. space agency. Sixty percent gave NASA a good to excellent rating. At the end of 1999, in a poll of about the same size, only 53% gave this high a rating. Was NASA being looked upon more favorably by the American public in 2005 than in 1999? Would you say that there is strong evidence to support your conclusion? (Always include all four steps in your test of significance.) [Source: Gallup, Public Favorable Toward NASA, Space Exploration, 2006, www.poll.gallup.com.]
E68. The legal drinking age in the United States is 21, and many people think that underage drinking is a serious problem. This question was asked in two recent separate polls: “Which one of the following do you think would have the most influence on persuading teenagers not to drink?” One poll randomly sampled 1008 adults, and the other randomly sampled 514 teens. Among the adults, 26% chose “parents” and 12% chose “friends.” Among the teens, 18% chose “parents” and 26% chose “friends.” (There was virtual agreement on the other choices, alcohol awareness programs and penalties.) 

[Source: USA Today teen drinking poll, 2001.]

a. Do adults choose “parents” significantly more often than teens choose “parents”? Would you say that there is strong evidence to support your conclusion?

b. Can you use the methods of this section to test for a significant difference between the proportion of teens who choose “parents” and the proportion of teenagers who choose “friends”? Why or why not?

E69. What is the trend in Internet use? National polls of about 1000 adults each estimated that 47% of adults logged on to the Internet for an hour or more daily in 2003 and that this increased to 51% by 2005. [Source: Gallup, Internet Catches More of Americans’ Time, 2006, www.poll.gallup.com.]

a. Is this a statistically significant increase?

b. Are you justified in saying that a majority of adults in the United States logged on to the Internet daily in 2005? Explain your reasoning.

E70. In the same polls on Internet use as referenced in E69, it was estimated that 33% of adults log on to the Internet for more than an hour daily but 27% never log on. Can you test to see if this is a significant difference using the methods of this section? Why or why not?

E71. You feed Diet A to a random sample of cows and Diet B to another random sample. Your null hypothesis is that the two diets produce the same proportion of contented cows. Your two-sample test has a $P$-value of 0.36. Select the one best interpretation of this $P$-value.

A. If the null hypothesis that there is no difference in the proportion of contented cows produced by the two diets is true, the probability is 0.36 that the proportion of contented cows produced by each diet won’t be very different.

B. You can’t conclude that the diets make a difference because it’s fairly likely to get as big a difference in the proportion of contented cows produced by the two diets as you did even if there is no difference at all in the overall proportion of contented cows that each diet produces.

C. The probability is 0.36 that the two diets differ in the proportion of contented cows that they produce. Because this probability is so high, you can reject the null hypothesis.

D. If the null hypothesis is true, then the probability is 0.36 that the two diets differ in the proportion of contented cows.

E. You can’t reject the null hypothesis because if, in fact, the null hypothesis is true, the probability is 0.36 that the difference in the true proportions is the same as in your experiment.

E72. Explain how knowledge of sampling distributions for differences in sample proportions is used in the development of the significance test of the difference of two proportions.
For E73 and E74: In this section, you have been testing the null hypothesis that $p_1 - p_2$ equals 0. The null hypothesis does not have to state that the difference in the two population proportions is 0. You can test for any difference you want. In the z-statistic for the next two problems, replace $p_1 - p_2$ with the difference in population proportions you are testing for rather than replacing it with 0. Then, when estimating the SE, \[ \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}, \] note that you should not use a pooled estimate of a common population proportion. The pooled estimate can be used only under the hypothesis that the two population proportions are equal, which is not what you are hypothesizing here.

E73. Return to the example on page 529, where a survey found that 113 of 1067 males were left-handed while 92 of 1170 females were left-handed. Perform a test to determine if there is statistically significant evidence that the percentage of males who are left-handed is greater than the percentage of females who are left-handed by more than 2 percentage points.

E74. You can take either Bus A or Bus B to school, and you’re concerned about being late to class. You check out Bus A on 100 randomly selected mornings and find that you would be late to class 5 mornings. You check out Bus B on 64 randomly selected mornings and find that you would be late 5 mornings. Because you live a bit closer to the stop for Bus B, you will use Bus A only if you can conclude that the difference in the proportions of late arrivals is more than 2%. Which bus will you use?

### 8.5 Inference for Experiments

An experiment uses the available subjects.

To use the methods for comparing two proportions discussed thus far—confidence interval and test of significance—the underlying condition was that the data came from random samples selected independently from two different populations. This is the typical situation with sample surveys, such as surveys comparing pet ownership rates for two different years. However, this is not the typical situation in experiments.

Experiments usually require investigators to use those subjects who volunteer or those experimental units available for the study. If there are two treatments, this group of subjects is then split into two groups by random assignment of the treatments. In such situations, the population is the group of subjects being used in the study.

Activity 8.5a will lead you through a simulation to help you see what happens to the sampling distribution of the difference of two proportions under this new method of randomization.

### ACTIVITY 8.5a

**Random Assignment in an Experiment**

**What you’ll need:** a group of students (the number in the group must be divisible by 4; any remaining students will be in charge of passing out the cards); a small card for each student, half marked “sit” and half marked “stand” (you can use 3×5 cards or old playing cards); a second small card for each student, 75% marked with a C and 25% marked with a B; a set of B/C cards for every student or small group of students, if you want to speed things up; paper or a blackboard at the front of the room for a large dot plot
Suppose your entire class is afflicted with terminal boredom. In this experiment, you will be randomly assigned one of two treatments, *sit* or *stand*, for your terminal boredom. You will also be dealt a card that shows the result of your treatment, B for *still bored*, or C for *cured*. The goal is to study the sampling distribution of the difference between the proportions cured by the two treatments.

1. Shuffle the sit/stand cards and give one to every student. Implement your assigned treatment by either sitting or standing.

2. Shuffle the C/B cards and give one to every student. This card tells you whether your treatment worked or not. If you get a card marked C, you are cured! If you get a card marked B, you are still terminally bored.

3. For each treatment, find the proportion of students who were cured. Then find the difference in the proportions who were cured, \( \hat{p}_{\text{sit}} - \hat{p}_{\text{stand}} \). Here \( \hat{p}_{\text{sit}} \) represents the proportion of sitting students who were cured, and \( \hat{p}_{\text{stand}} \) represents the proportion of standing students who were cured. Mark the difference on a dot plot.

4. Run this experiment again. Because it's obvious that this particular treatment has no effect whatsoever on terminal boredom, you might as well keep the same C or B card that you were given before. To see the effect of a different random assignment of treatments, shuffle the sit/stand cards and hand them out again. For each treatment, find the proportion of students who were cured. Then find the difference in the two proportions, \( \hat{p}_{\text{sit}} - \hat{p}_{\text{stand}} \). Mark this second difference on the dot plot.

5. Repeat step 4 until you have enough differences to enable you to see the shape, center, and spread of the sampling distribution.

Here's how to speed things up: Make duplicate sets of the B/C cards, with 75% marked C and 25% marked B. Shuffle and deal a set of the cards to make two equal piles, the one on the left representing the “sit” treatment and the one on the right representing the “stand” treatment. (You have just randomly assigned treatments to your subjects.) Compute \( \hat{p}_{\text{sit}} - \hat{p}_{\text{stand}} \) and record the difference as before on the dot plot.

6. Describe the shape, mean, and standard error of your approximate sampling distribution. Is your distribution centered where you expect it to be?

7. Compare the SE to what you get by using the formula for the standard error of the difference in sample proportions,

\[
\sigma_{\hat{p}_{1} - \hat{p}_{2}} = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}
\]

For this experiment, \( n_1 \) and \( n_2 \) are each half the number of students, and \( p_1 \) and \( p_2 \) are each 0.75, the probability of being cured. What can you conclude?

Display 8.28 shows the sampling distribution from Activity 8.5a for a class of 40 students. The students participated in a sham experiment in which the
treatments were of no value whatsoever in curing their affliction. The histogram shows the difference in the proportion of students “cured” by the two sham treatments. This class used a computer to generate 10,000 rerandomizations for the different “treatments” and thus has 10,000 differences.

Display 8.28 The difference in the proportion of students “cured” by two sham treatments when the probability of a cure is 0.75 no matter which treatment the student received. Here \( n_1 = n_2 = 20 \).

Note that this approximate sampling distribution of the differences in proportions from an experiment

- looks approximately normal
- is centered at 0, as expected
- has standard error about 0.137

Now comes the great piece of luck. You don’t have to do a simulation to get the standard error. The formula you used for a sample survey works fine:

\[
SE = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}} = \sqrt{\frac{0.75(1 - 0.75)}{20} + \frac{0.75(1 - 0.75)}{20}} = 0.137
\]

Of course, in the real-life situations you will be studying next, you will not know the true success rates of the treatments and will have to estimate them from the proportions of successes in the treatment groups. But that is nothing new—this formula will continue to work fine just as you have used it before.

**Confidence Interval for a Difference in Proportions from an Experiment**

For experiments, the form of the confidence interval and all the computations remain the same, as do the conditions on sample size. But the interpretation is a little different because you are estimating this difference:

\[
\text{proportion of successes if all units had received Treatment A} - \text{proportion of successes if all units had received Treatment B}
\]
Confidence Interval for a Difference in Proportions from an Experiment

The confidence interval for the difference \( p_1 - p_2 \) in the proportions of successes in a population of experimental units subjected to two different treatments is

\[
(p_1 - p_2) \pm z^* \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}
\]

Here \( \hat{p}_1 \) is the observed proportion of successes in the group of size \( n_1 \) given the first treatment, and \( \hat{p}_2 \) is the observed proportion of successes in the group of size \( n_2 \) given the second treatment. (The group sizes do not have to be equal.) The conditions that must be met in order to use this formula are that

- the two treatments are randomly assigned to the population of available experimental units
- \( n_1 \hat{p}_1, n_1(1 - \hat{p}_1), n_2 \hat{p}_2, \) and \( n_2(1 - \hat{p}_2) \) are all at least 5

Example: Confidence Interval for an Experiment

From 1986 to 1988, an important study on the treatment of patients with AIDS-related complex (ARC) was run by a collaborative group of researchers from eight European countries and Australia. (AIDS-related complex is a condition that generally leads to a diagnosis of AIDS.) A total of 134 patients with ARC agreed to participate in the study, which was a double-blind, randomized clinical trial. Two treatments were to be compared. Each patient was given either zidovudine (commonly known as AZT) by itself or a combination of AZT and acyclovir (ACV). One outcome measure was the number of ARC patients who developed AIDS during the 1 year of the study (Display 8.29).

<table>
<thead>
<tr>
<th>Treated with</th>
<th>AZT</th>
<th>AZT + ACV</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Developed AIDS?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>12</td>
<td>10</td>
<td>22</td>
</tr>
<tr>
<td>No</td>
<td>55</td>
<td>57</td>
<td>112</td>
</tr>
<tr>
<td>Total</td>
<td>67</td>
<td>67</td>
<td>134</td>
</tr>
</tbody>
</table>

Display 8.29  ARC patients progressing to AIDS during the study.


The group of ARC patients weren’t randomly selected from the population of all ARC patients but were volunteers who had to give “informed consent.” Thus, they might be quite different from the population of all ARC patients and should not be regarded as a random sample from any specific population.
However, the treatments were randomly assigned to the ARC patients who were available. (This type of randomization is typical in medical research, agricultural research, and many other kinds of experiments for which getting a simple random sample from the entire population of interest is impossible.)

During the life of the study, 12 of 67 patients (or 17.9%) using AZT alone progressed to AIDS, and 10 of 67 patients (or 14.9%) using AZT plus ACV progressed to AIDS. Find a 95% confidence interval for the difference between \( p_1 \), the proportion of patients in this study who would have developed AIDS if they all had taken the AZT treatment, and \( p_2 \), the proportion who would have developed AIDS if they all had taken the AZT plus ACV treatment.

**Solution**

The first condition that needs to be met is that the treatments were randomly assigned to the subjects. The second condition is that \( n_1 \hat{p}_1 = 12, n_1(1 - \hat{p}_1) = 55, n_2 \hat{p}_2 = 10, \) and \( n_2(1 - \hat{p}_2) = 57 \) are all 5 or more. Both conditions are met here.

The 95% confidence interval for the difference between the two proportions \( p_1 \) and \( p_2 \) is

\[
(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}
\]

\[
= (0.179 - 0.149) \pm 1.96 \sqrt{\frac{(0.179)(0.821)}{67} + \frac{(0.149)(0.851)}{67}}
\]

\[
= 0.03 \pm 0.125
\]

You also can write this confidence interval as \((-0.095, 0.155)\).

If all 134 patients could have been given each treatment, we estimate that the difference between the rate of progression to AIDS if all had been given AZT alone and the rate of progression if all had been given AZT plus ACV would be someplace between a 9.5% difference in favor of AZT alone and a 15.5% difference in favor of AZT plus ACV. Note that a difference of 0 lies well within the confidence interval. This means that if the difference in the proportions of patients who would progress to AIDS is actually 0, getting a difference of 0.03 in our treatment groups is reasonably likely. Thus, we aren’t convinced that the two therapies differ with respect to the rate of progression to AIDS in this group of patients.

### DISCUSSION

**Confidence Interval for a Difference in Proportions from an Experiment**

D46. Discuss how the randomization used in a two-population sample survey differs from the randomization used in a two-treatment experiment.

D47. The ARC data in the example came from a “double-blind, randomized clinical trial.” What is the meaning of each part of this phrase?

D48. Suppose the difference in the proportion of patients who improve with Treatment A and the proportion of patients who improve with Treatment B is 0.05. Does this mean that the majority of patients improve with Treatment A? Does it mean that Treatment B isn’t very good?
Significance Test for a Difference in Proportions from an Experiment

You now know that the sampling distribution of the difference between two proportions will still be approximately normal under the randomization scheme used in experiments and that the formula for the estimated standard error remains the same. Thus, for a significance test you can construct a test statistic just as you did in the case of sample survey data. The main difference between the two testing scenarios is in the stating of the hypotheses and conclusions. An outline of a significance test for comparing proportions from an experiment is given in the box.

Components of a Significance Test for the Difference of Two Proportions from an Experiment

1. Check conditions.
   Two treatments, A and B, are randomly assigned to available experimental units, with \( n_1 \) receiving Treatment A and \( n_2 \) receiving Treatment B. With \( \hat{p}_1 \) representing the proportion of successes from Treatment A and \( \hat{p}_2 \) representing the proportion of successes from Treatment B, all of the following values must be at least 5:
   \[
   n_1 \hat{p}_1, \quad n_1(1 - \hat{p}_1), \quad n_2 \hat{p}_2, \quad n_2(1 - \hat{p}_2)
   \]

2. Write a null and an alternative hypothesis.
   The null hypothesis is that the two treatments would have resulted in the same proportion of successes if each treatment could have been given to all subjects. The alternative hypothesis for a two-sided test is that the two treatments would not have resulted in the same proportion of successes if each treatment could have been given to all subjects. The alternative hypothesis for a one-sided test is that a specific treatment would have resulted in a higher proportion of successes than the other if each treatment could have been given to all subjects.

3. Compute the test statistic and a \( P \)-value.
   \[
   z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}
   \]
   where \( \hat{p} = \frac{\text{total number of successes in both groups}}{n_1 + n_2} \)

The \( P \)-value is the probability of getting a value of \( z \) as extreme as or even more extreme than that from your samples if \( H_0 \) is true. See Display 8.30.

(continued)
4. Write a conclusion.
State whether you reject or do not reject the null hypothesis. Link this
collection to your computations by comparing \( z \) to the critical value, \( z^* \), or
by appealing to the \( P \)-value. You reject the null hypothesis if \( z \) is more extreme
than \( z^* \) or, equivalently, if the \( P \)-value is smaller than the level of significance,
\( \alpha \). Then write a sentence giving your conclusion in the context of the situation.

To see how to apply these ideas to an experiment, return to the clinical trial
experiment comparing the two treatments for AIDS-related complex (ARC).

**Example: Significance Test for an Experiment**

The investigators’ next question dealt with the survival rate among 131 patients
who had already developed AIDS before entering the study. These data are shown
in Display 8.31.

<table>
<thead>
<tr>
<th>Survived?</th>
<th>Treated with</th>
<th>AzT</th>
<th>AzT + ACV</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td></td>
<td>28</td>
<td>13</td>
<td>41</td>
</tr>
<tr>
<td>Yes</td>
<td></td>
<td>41</td>
<td>49</td>
<td>90</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>69</td>
<td>62</td>
<td>131</td>
</tr>
</tbody>
</table>

Display 8.31 Treatment results among AIDS patients.

Here \( \hat{p}_1 = \frac{41}{69} \approx 0.594 \), where \( \hat{p}_1 \) is the proportion who survived in the group
treated with AzT alone, and \( \hat{p}_2 = \frac{49}{62} \approx 0.790 \), where \( \hat{p}_2 \) is the proportion who
survived in the group treated with AzT plus ACV. In order for a new therapy (in
this case, AzT plus ACV) to become accepted practice among physicians, the
research must show that the new therapy is an improvement over the old one.
Can that claim be made based on these data?

**Solution**

Because the researchers are attempting to show that the new therapy is better than
the old one, a one-sided significance test is in order. The hypotheses are

\( H_0: \) The new therapy is not better than the old, or \( p_1 = p_2 \), where \( p_1 \) is the
proportion of patients who would have survived if all 131 patients could have been given AzT alone and \( p_2 \) is the proportion of patients who would have survived if they all could have been given AzT plus ACV.

\( H_1: \) The new therapy is better than the old one, or \( p_1 < p_2 \).
Treatments were randomly assigned to the available subjects, and each of
\[ n_1 \hat{p}_1 = 41 \quad n_1(1 - \hat{p}_1) = 28 \quad n_2 \hat{p}_2 = 49 \quad n_2(1 - \hat{p}_2) = 13 \]
is at least 5.

The pooled estimate, \( \hat{p} \), of the probability that a patient will survive, under the null hypothesis that the treatment a patient received made no difference in whether he or she survived, is \( \frac{90}{131} \), or about 0.687. The test statistic then takes on the value

\[ z = \frac{\text{statistic} - \text{parameter}}{\text{standard deviation of statistic}} \]
\[ = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \]
\[ = \frac{(0.594 - 0.790) - 0}{\sqrt{0.687(1 - 0.687)\left(\frac{1}{69} + \frac{1}{62}\right)}} = -2.415 \]

This test statistic is based on the difference of two approximately normally distributed random variables, \( \hat{p}_1 \) and \( \hat{p}_2 \). You learned in Chapter 7 that such a difference is itself approximately normal. Thus, you can use Table A on page 824 to find the \( P \)-value. The test statistic \( z = -2.42 \) has a (one-sided) \( P \)-value of 0.0078, as illustrated in Display 8.32.

Display 8.32 \( P \)-value for a one-sided test with \( z = -2.42 \).

This \( P \)-value is very small, so reject the null hypothesis. If all subjects in the experiment could have been given the AZT plus ACV treatment, you are confident that there would have been a larger survival rate than if all had received only AZT. The difference between the survival rates for these two treatments is too large to be attributed to chance variation alone.

---

**DISCUSSION**

**Significance Test for a Difference in Proportions from an Experiment**

D49. Discuss the differences in stating hypotheses for testing the equality of proportions in a two-population sample survey versus a two-treatment experiment.

D50. In the AZT versus AZT plus ACV example, the null hypothesis was rejected. Which type of error might have been made in the significance test? Discuss the practical ramifications of making this error in this context.
Inference for an Observational Study

Sometimes it is impossible, either ethically or practically, to assign treatments to subjects or experimental units. This often is the situation in epidemiological studies of the spread or progression of a disease, because it is unethical to give a subject a harmful disease as part of an experiment. However, possible associations between living habits and the progression of a disease may be observed by studying many people with the disease to see if, for example, people who exercise regularly have slower progression of the disease than people who don’t exercise regularly. As you saw in Chapter 4, studies in which the conditions of interest are already built into the units being studied are called observational studies. Even though no randomization is involved, inference procedures can still be applied if you are very careful in stating the question and your conclusion. What you are asking is, “Could the result I see in the observed data have reasonably happened by chance?” In other words, if the data had been generated through a randomized experiment, would such a result be likely? If the answer is no, then there is evidence of an association that should be investigated further. Remember, observational studies do not allow you to make cause-and-effect conclusions or to estimate population parameters in a rigorous way, but they can provide evidence of possible associations.

Example: Confidence Interval in an Observational Study

Do you know someone who is trying to quit smoking? Does the person have a plan to stop? How successful has he or she been? Some have argued that a spur-of-the-moment decision to stop smoking (cold turkey, if you will) is just as successful as the best-laid plan to stop smoking. In a recent study on this question, researchers observed 611 documented attempts to quit smoking. Among the 297 unplanned, spur-of-the-moment attempts, 194 succeeded 6 months or longer. Among the 314 planned attempts, 133 succeeded 6 months or longer. [Source: www.medpagetoday.com.] Use a 95% confidence interval to explore a possible association between unplanned attempts and success rate.

Solution

The conditions for a standard confidence interval for the difference between proportions are not met because there was no randomization in this study. The researchers studied the 611 attempts that were available to them from data collected in a national survey, but the “treatments,” planned and unplanned attempts to quit smoking, were not randomly assigned. Nevertheless, the 95% confidence interval can be informative. This develops from standard procedures:

\[
\hat{p}_1 = \frac{194}{297} = 0.65, \quad \hat{p}_2 = \frac{133}{314} \approx 0.42
\]

\[
(\hat{p}_1 - \hat{p}_2) \pm z^* \cdot \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}
\]

\[
= (0.65 - 0.42) \pm 1.96 \cdot \sqrt{\frac{0.65(0.35)}{297} + \frac{0.42(0.58)}{314}} = 0.23 \pm 0.08
\]

The confidence interval is (0.15, 0.31).
This confidence interval provides some evidence that there is an association between the type of attempt and the success of the attempt. Those who attempt to quit smoking spur-of-the-moment might have a success rate that is larger than those who use planned attempts by, perhaps, something between 15% and 31%. However, this study cannot determine whether the method itself resulted in a higher success rate or some other variable was responsible, such as a difference in the willpower of people who choose the two methods. Nor can this study determine if the difference applies to the larger population of smokers who attempt to quit. All that can be concluded is that this difference cannot easily be attributed to chance alone. Thus, this result might suggest to the researchers that they should collect more data on this question, although they probably can never do a randomized study. (They could not randomly assign a person to make an immediate decision to stop smoking.)

Example: Significance Test in an Observational Study

Does a helmet law have much impact on whether bicycle riders actually wear helmets? Anticipating the passing of a helmet law, the government of North Carolina commissioned a study of helmet usage among bicycle riders at various locations across the state in 1999. This study provided baseline data for studying the effects of the law. In 2002, 6 months after the law was passed, a similar study was conducted at the same types of locations. Some of the resulting data summaries are shown in Display 8.33.

<table>
<thead>
<tr>
<th>Location Type</th>
<th>% Helmet Use 1999</th>
<th>% Helmet Use 2002</th>
<th>Sample Size 1999</th>
<th>Sample Size 2002</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local streets</td>
<td>16</td>
<td>19</td>
<td>1116</td>
<td>848</td>
</tr>
<tr>
<td>Collector streets</td>
<td>25</td>
<td>30</td>
<td>592</td>
<td>513</td>
</tr>
<tr>
<td>Greenways</td>
<td>42</td>
<td>55</td>
<td>404</td>
<td>369</td>
</tr>
<tr>
<td>Mountain biking trails</td>
<td>84</td>
<td>90</td>
<td>336</td>
<td>219</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>2448</td>
<td>1949</td>
</tr>
</tbody>
</table>

Display 8.33 Helmet usage at various locations. [Source: University of North Carolina Highway Safety Research Center, www.hsrc.unc.edu.]

Is there sufficient evidence to say that the difference in the percentage using helmets on local streets before the law and after the law is so large that it cannot reasonably be attributed to chance alone?

Solution

Even though some of the sites for observing riders were randomly selected, the riders observed were those who just happened to be there at the time of observation. In addition, there were no randomly assigned treatments and no controls on any of the myriad other variables that could affect helmet use. Nevertheless, a test of significance can shed some light on the possible association between the law and helmet usage. The null hypothesis is that there is no difference in the proportions of helmet users on local streets for 1999 and 2002. The alternative hypothesis of interest is that the 2002 rate of helmet use is higher than the 1999 rate (perhaps because of the law).
Under the null hypothesis of no difference in proportions, the pooled estimate of helmet use on local streets is

\[
\hat{p} = \frac{(0.16)(1116) + (0.19)(848)}{1116 + 848} = 0.173
\]

If you let \( \hat{p}_1 = 0.16 \) denote the proportion of helmet users observed on local streets in 1999 and \( \hat{p}_2 = 0.19 \) denote the proportion observed in 2002, the test statistic becomes

\[
z = \frac{0.16 - 0.19}{\sqrt{(0.173)(1 - 0.173)\left(\frac{1}{1116} + \frac{1}{848}\right)}} = -1.741
\]

If this were a formal one-sided test of significance, the \( P \)-value would be 0.04. This indicates that, if there was no change in the percentage of helmet users, there would be very little chance of seeing a difference this large in the percentage wearing helmets in samples of randomly selected bicycle riders. Thus, there is some evidence that there was a higher rate of helmet-wearing in 2002, but that could be due to a combination of many factors, one of which might be the law. The law could have coincided with an increased emphasis on safety education, a general trend in the behavior of bicycle riders, the age or experience of the riders, or even a change in the weather. None of these possible factors are measured, controlled, or balanced by randomization.

**Inference for an Observational Study**

D51. Discuss the differences in the types of conclusions that can be reached in a randomized experiment versus those that can be reached in an observational study.

D52. There is a large controversy in educational research about whether randomized experiments can be used to reach conclusions about curriculum materials and methods. For example, suppose a new method for teaching fractions is to be compared to a standard method that has been in use for years. In order to be accepted as “better,” the new method must be tested against the old in real classroom situations. Should these tests be set up as randomized experiments or as observational studies? What do you suppose is a key argument in favor of randomized experiments? What do you suppose is a key argument against trying to conduct randomized experiments?

**Summary 8.5: Inference for Experiments**

Both confidence interval estimation and tests of significance proceed in inference for experiments much as they did in inference for sample surveys, but the hypotheses and conclusions must be stated differently. This is because an experiment generally does not result in a comparison based on independent random samples from two populations. Rather, one population of available experimental units is split randomly into two groups by the assignment of treatments. As long as the basic sample-size considerations are met, however, the same methods used in the analysis of sample surveys apply.
Chapter 8  Inference for Proportions

Confi dence Interval for a Diff erence in Proportions from an Experiment

P52. Th e landmark Physicians’ Health Study study of the eff ect of low-dose aspirin on the incidence of heart attacks (see E79 on page 552) checked to see if aspirin use was associated with increased risk of ulcers. Th is was a randomized, double-blind, placebo-controlled clinical trial. Male physician volunteers with no previous important health problems were randomly assigned to an experimental group \((n_1 = 11,037)\) or a control group \((n_2 = 11,034)\). Those in the experimental group were asked to take a pill containing 325 mg of aspirin every second day, and those in the control group were asked to take a placebo pill every second day. Of those who took the aspirin, 169 got an ulcer, compared to 138 in the placebo group. Find and interpret a 95% confi dence interval for the diff erence of two proportions. [Source: Steering Committee of the Physicians’ Health Study Research Group, “The Final Report on the Aspirin Component of the Ongoing Physicians’ Health Study,” New England Journal of Medicine 321, no. 3 (1989): 129–35; www.content.nejm.org.]

P53. In a classic 1962 social experiment in Ypsilanti, Michigan, 3- and 4-year-old children were assigned at random to either a treatment group receiving two years of preschool instruction or a control group receiving no preschool instruction. Th e participants in the study were followed through their teenage years. One variable recorded was the number in each group arrested for a crime by the time they were 19 years old. Th e data are shown in Display 8.34. [Source: F. Ramsey and D. Schafer, The Statistical Sleuth (Duxbury, 2002), p. 546; Time, July 29, 1991.]


<table>
<thead>
<tr>
<th>Preschool</th>
<th>Arrested</th>
<th>Not Arrested</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>19</td>
<td>42</td>
</tr>
<tr>
<td>No Preschool</td>
<td>32</td>
<td>30</td>
</tr>
</tbody>
</table>

a. Are the conditions met for constructing a confidence interval for the difference of two proportions?

b. Estimate the true difference between the proportions arrested, using a 90% confidence interval.

c. Carefully state a conclusion in the context of the problem.

d. Th is is an experiment, but is a clear cause-and-effect relationship established here? Explain your reasoning.

Significance Test for a Diff erence in Proportions from an Experiment

P54. Th e battle against high cholesterol has been raging for years. A randomized clinical trial compared two drugs designed to lower cholesterol (Lipitor and Zocor) with regard to both the cholesterol outcome and the rate of heart attacks over a 5-year period. It is the prevention of heart attacks that is the ultimate goal, and Zocor is much less expensive than Lipitor (because Zocor’s patent expired in 2006). High doses of Lipitor won out over low doses of Zocor on lowering cholesterol, but the heart attack issue was much closer. Of 4420 patients on Lipitor, 411 had heart attacks. Of 4452 patients on Zocor, 463 had heart attacks. [Source: USA Today, November 16, 2005.]

a. Is this evidence of a significant diff erence in heart attack rates for the two drugs?

b. Prior to the study, many experts thought that the high doses of Lipitor should be much more eff ective against heart attacks than the low doses of Zocor. Conduct the appropriate one-sided test to check this claim. What is your conclusion? How do you reconcile it with your conclusion in part a?

P55. In P54, which type of error could have been made in part a? In part b? Which do you think is the more serious error in the practical context of this problem? Explain your reasoning.
Inference for an Observational Study

P56. Refer to the example on the attempts to quit smoking on page 547. Set up a statistical analysis to test the claim that the spur-of-the-moment method is actually better than the planned method for quitting. State your conclusion carefully in light of the fact that this is an observational study.

P57. Do victims of violence exhibit more violence toward others? In a classic study of this question, a researcher found records of 908 people who had been abused as children. Then, matching the demographic characteristics of this group as closely as she could, she found records of 667 individuals who had not been abused as children. Based on further searches of records, she discovered that 103 of the abused children later perpetrated violent crimes, while 53 of the children not abused perpetrated violent crimes. [Source: C. S. Wisdom, “The Cycle of Violence,” Science 244 (1989): 160.]

a. What kind of study was this: sample survey, experiment, or observational study?

b. Estimate the difference in violent crime rates for the two groups, using a 90% confidence interval.

c. Write a conclusion in the context of the problem, being careful to consider the type of design that was used. Does this prove that abuse of children leads to violent crime down the road?

P58. Refer to the example on helmet use in North Carolina on page 548. Among the other three locations (collector streets, greenways, and mountain biking trails), which shows the strongest association between helmet use and the law?

Exercises

E75. A famous medical experiment was conducted by Nobel laureate Linus Pauling (1901–1994), who believed that vitamin C prevents colds. His subjects were 279 French skiers who were randomly assigned to receive vitamin C or a placebo. Of the 139 given vitamin C, 17 got a cold. Of the 140 given the placebo, 31 got a cold. [Source: L. Pauling, “The Significance of the Evidence About Ascorbic Acid and the Common Cold,” Proceedings of the National Academy of Sciences 68 (1971): 2678–81.]

a. Find and interpret a 95% confidence interval for the difference of two proportions.

b. Find a 99% confidence interval for the difference of two proportions. Does your conclusion change from your interpretation in part a?

c. Find and interpret a 95% confidence interval for the difference of two proportions.

d. Find a 99% confidence interval for the difference of two proportions. Does your conclusion change from your interpretation in part a?

E76. A randomized clinical trial on Linus Pauling’s claim that vitamin C helps prevent the common cold was carried out in Canada among 818 volunteers, with results reported in 1972. The data showed that 335 of the 411 in the placebo group got colds over the winter in which the study was conducted, while 302 of the 407 in the vitamin C group got colds. [Source: T. W. Anderson, D. B. Reid, and G. H. Beaton, “Vitamin C and the Common Cold,” Canadian Medical Association Journal 107 (1972): 503–8.]

a. Find and interpret a 95% confidence interval for the difference of two proportions.

b. Find a 99% confidence interval for the difference of two proportions. Does your conclusion change from your interpretation in part a?

c. Find and interpret a 95% confidence interval for the difference of two proportions.

d. Find a 99% confidence interval for the difference of two proportions. Does your conclusion change from your interpretation in part a?

E77. In 1954, the largest medical experiment of all time was carried out to test whether the newly developed Salk vaccine was effective in preventing polio. This study incorporated all three characteristics of an experiment: use of a control group of children who received a placebo injection (an injection that felt like a regular immunization but contained only salt water), random assignment of children to either the placebo injection group or the Salk vaccine injection group, and assignment of each treatment to several hundred thousand children. Of the 200,745 children who received the Salk vaccine,

E78. Two different methods of treating wool with a mothproofing agent were tested in a laboratory by randomly dividing 40 wool samples into two groups of 20, one group randomly assigned to Method A and the other to Method B. Moth larvae were then attached to each sample of wool and observed for a fixed period of time. For Method A, 8 of the 20 larvae died; for Method B, 12 of the 20 larvae died. Is there evidence of a difference in the effectiveness of these two methods? Answer the question by constructing and interpreting an appropriate 95% confidence interval. [Source: G. E. P. Box, W. G. Hunter, and J. S. Hunter, Statistics for Experimenters (Wiley, 1978), pp. 132–33.]

E79. The Physicians’ Health Study conducted a famous experiment on the effect of low-dose aspirin on heart attacks. This was a randomized, double-blind, placebo-controlled clinical trial. Male physician volunteers with no previous important health problems were randomly assigned to an experimental group \((n_1 = 11,037)\) or a control group \((n_2 = 11,034)\). Those in the experimental group were asked to take a pill containing 325 mg of aspirin every second day, and those in the control group were asked to take a placebo pill every second day. After about 5 years, there were 139 heart attacks in the aspirin group and 239 in the placebo group. Is there significant evidence that aspirin use reduces the rate of heart attacks in this group? Test at the 1% level. [Source: Steering Committee of the Physicians’ Health Study Research Group, “The Final Report on the Aspirin Component of the Ongoing Physicians’ Health Study,” New England Journal of Medicine 321, no. 3 (1989): 129–35; www.content.nejm.org.]

E80. A related part of the landmark study discussed in E79 dealt with the question of whether aspirin reduces the seriousness of a heart attack should one occur. Of the 139 volunteers in the aspirin group who had a heart attack, 10 died. Of the 239 in the placebo group who had a heart attack, 26 died. Is the difference in death rate statistically significant?

E81. “Exercise Helps Delay Onset of Dementia” read a headline in the Gainesville Sun on January 17, 2006. Researchers followed a group of people age 65 and older from 1994 to 2003. Among the 1185 free of dementia at the end of this time, 77% reported exercising three or more times a week. Among the 158 who showed signs of dementia at the end of the period, 67% reported exercising three or more times a week.

a. Is this an experiment, a sample survey, or an observational study?

b. Test to see if the difference in these proportions is statistically significant. Then state your conclusions carefully, in light of your answer to part a. Can this study demonstrate that exercise causes a delay in the onset of dementia?

E82. In one early study (1956) of the relationship between smoking and mortality, Canadian war pensioners were asked about their smoking habits and then followed to see if they were alive 6 years later. Of the 1067 nonsmokers, 950 were still alive. Of the 402 pipe smokers, 348 were still alive. [Source: G. W. Snedecor and W. G. Cochran, Statistical Methods, 6th ed. (Iowa State Press, 1967), pp. 215–16.]

a. Is this an experiment, a sample survey, or an observational study?

b. Is there statistically significant evidence that pipe smoking increases the rate of mortality? State your conclusions carefully, in light of your answer to part a.
E83. The great baseball player Reggie Jackson batted .262 during regular season play and .357 during World Series play. He was at bat 9864 times in the regular season and 98 times in the World Series. Can this difference reasonably be attributed to chance, or did Reggie earn his nickname “Mr. October”? (The World Series is played in October.) Use $\alpha = 0.05$. Note that you do not have two random samples from different populations. You have Reggie’s entire record, so all you can decide is whether the difference is about the size you would expect from chance variation or you should look for some other explanation. [Source: Major League Baseball, February 23, 2002, www.mlb.com]

E84. At Bunker Hill, near Boston, in one of the first battles of the American Revolution, 2400 British troops were engaged, of which 1054 were wounded. Of the 1054 wounded, 226 died. Out of 1500 American participants in the battle, losses were estimated at 140 killed and an additional 271 wounded who didn’t die of their wounds. (The number of casualties varies somewhat depending on the source.) You have all the necessary information about the two populations. Can the difference in the pairs of proportions in I and II reasonably be attributed to chance variation, or should you look for another explanation? Use $\alpha = 0.05$. [Source: Christopher Ward, The War of the Revolution (New York: Macmillan, 1952), p. 96.]

I. the proportion of American troops who were wounded (including those killed) and the proportion of the British troops who were wounded (including those killed)

II. the proportion of the American troops who were killed and the proportion of the British troops who were killed

E85. A test of significance asks the question “Could the result I see have happened by chance, or should I look for another cause?” In randomized experiments, the answer to this question can be found simply by studying what happens in repeated randomizations, without recourse to standard errors, test statistics, or the normal distribution. Here is the construction of such a test based on the AIDS data in Display 8.31 and the underlying null hypothesis that there is no difference in the death rates for the two treatments, AZT and AZT plus ACV.

1. Make 90 chips numbered 1 to represent those alive and 41 chips numbered 0 to represent those not alive at the end of the study time.

2. Randomly split the 131 chips into two groups, one with 69 chips (representing those treated with AZT alone) and one with 62 chips (representing those treated with AZT plus ACV).

3. Calculate the difference between the two proportions of 1’s (subtracting AZT plus ACV from AZT).

4. Repeat the process many times and plot the values of the generated differences in proportions. Under the null hypothesis, each of the generated differences has the same chance of occurring.

A plot of 200 runs of this simulation procedure is shown in Display 8.35 on the next page.
DiffProp

Measures from Sample of AZT data

Dot Plot

Display 8.35 A plot of 200 runs of a randomization process.

a. The observed difference in the proportions for the data from the real experiment was about $-0.20$. How many times, out of 200, did the randomization procedure generate a difference this small or smaller? What does this suggest about the chance of seeing a value this small under the null hypothesis?

b. Explain what your conclusion would be regarding the two treatments based solely on the randomization procedure. Does your conclusion agree with the one reached in the example on pages 545–546?

E86. Refer to the data in P53 on page 550. Use the randomization procedure from E85 to test the hypothesis that preschool makes no difference in whether a person is arrested for a violent crime.

E87. Refer to the data in the example on page 542 (Display 8.29). Use the randomization procedure from E85 to test the hypothesis that AZT plus ACV is better than AZT alone for reducing the incidence of AIDS. Does your conclusion agree with the one reached in the example by constructing and interpreting a confidence interval?

Chapter Summary

To use the confidence intervals and significance tests of this chapter, you need either a random sample from a population that is made up of “successes” and “failures” or two random samples taken independently from two distinct such populations. If the study is an experiment, you should have two treatments randomly assigned to the available subjects. (When there is no randomness involved, you proceed with the test only if you state clearly the limitations of what you have done. If you reject the null hypothesis in such a case, all you can conclude is that something happened that can’t reasonably be attributed to chance.)

You use a confidence interval if you want to find a range of plausible values for

- $p$, the proportion of successes in a population
- $p_1 - p_2$, the difference between the proportion of successes in one population and the proportion of successes in another population

Both confidence intervals you have studied have the same form:

$$\text{statistic} \pm (\text{critical value}) \cdot (\text{standard deviation of statistic})$$

Both significance tests you have studied include the same steps:

1. Justify your reasons for choosing this particular test. Discuss whether the conditions are met, and decide whether it is okay to proceed if they are not strictly met.
2. State the null hypothesis and the alternative hypothesis.
E88. A 6th-grade student, Emily Rosa, performed an experiment to test the validity of “therapeutic touch.” According to an article written by her RN mother, a statistician, and a physician: “Therapeutic Touch (TT) is a widely used nursing practice rooted in mysticism but alleged to have a scientific basis. Practitioners of TT claim to treat many medical conditions by using their hands to manipulate a ‘human energy field’ perceptible above the patient’s skin.” To investigate whether TT practitioners actually can perceive a “human energy field,” 21 experienced TT practitioners were tested to determine whether they could correctly identify which of their hands was closest to the investigator’s hand. They placed their hands through holes in a screen so they could not see the investigator. Placement of the investigator’s hand was determined by flipping a coin. Practitioners of TT identified the correct hand in only 123 of 280 trials. [Source: Journal of the American Medical Association 279 (1998): 1005–10.]

a. What does “blinded” mean in the context of this experiment? How might it have been done? Why wasn’t double-blinding necessary?

b. Write appropriate null and alternative hypotheses for a significance test.

c. Is it necessary to actually carry out the test, based on these data?

d. The article says, “The statistical power of this experiment was sufficient to conclude that if TT practitioners could reliably detect a human energy field, the study would have demonstrated this.” What does this sentence mean?

E89. “Most teens are not careful enough about the information they give out about themselves online.” In a survey that randomly sampled 971 teenagers who have online access, 78% agreed with this statement. [Source: Pew Internet and American Life Project, 2005, www.pewinternet.org.]

a. Compute the 95% confidence interval for the population proportion of teenagers who would agree with the statement.

b. Interpret this confidence interval, making it clear exactly what it is that you are 95% sure is in the confidence interval.

c. Explain the meaning of 95% confidence.

E90. A 2001 Gallup poll found that 51% of the American public assigned a grade of A or B to the public schools in their community. In 2000, the comparable figure was 47%. Assuming a sample size of 1108 in both 2000 and 2001, find and interpret a 90% confidence interval for the difference of two proportions. [Source: Gallup, www.gallup.com.]

E91. A study of all injuries from the two winter seasons 1999–2000 and 2000–2001 at the three largest ski areas in Scotland found that of the 531 snowboarders who were injured, 148 had fractures. Of the 952 skiers who were injured, 146 had fractures. (For both groups, most of the other injuries were sprains, lacerations, or bruising.) [Source: www.ski-injury.com, March 29, 2002.]

a. Is this difference statistically significant at the 0.05 level?

b. There are about twice as many skiers as snowboarders. Can you use this fact and the data from this study to determine whether snowboarders are more likely than skiers to be injured?
E92. To study the brain’s response to placebos, researchers at UCLA gave 51 patients with depression either an antidepressant or a placebo. The article reports that the “51 subjects then were randomly assigned to receive 8 weeks of double-blind treatment with either placebo or the active medication. . . . Overall, 52% of the subjects (13 of 25) receiving antidepressant medication responded to treatment, and 38% of those receiving placebo (10 of 26) responded.” Perform a test to determine if the difference in the proportion who responded is statistically significant. Use $\alpha = 0.02$. [Source: A. F. Leuchter, I. A. Cook, E. A. Witte, M. Morgan, and M. Abrams, “Changes in Brain Function of Depressed Subjects During Treatment with Placebo,” *American Journal of Psychiatry* 159, no. 1 (January 2002): 122–29.]

E93. Suppose a null hypothesis is tested at the 0.05 level of significance in 100 different studies in which the null hypothesis is true.

a. How many Type I errors do you expect?

b. Find the probability that no Type I error is made in any of the 100 studies.

E94. To become a lawyer, a person must pass the bar exam for his or her state. Suppose you are searching for possible inequities that result in unequal proportions of males and females passing the bar exam. Each state has its own exam, so you check a random sample of the exam records in each of the 50 states. For each state, you plan to perform a significance test of the difference of the proportion of males and females passing the exam, using $\alpha = 0.05$. If the difference is statistically significant for any state, you will conclude that there has been some inequity. Do you see anything wrong with this plan?

E95. Suppose you use the formula for a 90% confidence interval for many different random samples from the same population. What fraction of times should the intervals produced actually capture the population proportion? Explain your reasoning.

E96. What is the best explanation of the use of the term “95% confidence”?

A. We can never be 100% confident in statistics; we can only be 95% confident.

B. The sample proportion will fall in 95% of the confidence intervals we construct.

C. In the long run, the population proportion will fall in 95% of the confidence intervals we construct.

D. We are 95% confident that the range of reasonably likely outcomes contains the population proportion.

E. We are 95% confident that we have made a correct decision when we reject the null hypothesis.

E97. How could this explanation of margin of error be improved?

Statisticians over the years have developed quite specific ways of measuring the accuracy of samples—so long as the fundamental principle of equal probability of selection is adhered to when the sample is drawn. For example, with a sample size of 1,000 national adults (derived using careful random selection procedures), the results are highly likely to be accurate within a margin of error of plus or minus three percentage points.

So, if we find in a given poll that President Clinton’s approval rating is 50%, the margin of error indicates that the true rating is very likely to be between 47% and 53%. It is very unlikely to be higher or lower than that. To be more specific, the laws of probability say that if we were to conduct the same survey 100 times, asking people in each survey to rate the job Bill Clinton is doing as president, in 95 out of those 100 polls, we would find his rating to be between 47% and 53%. In only five of those surveys would we expect his rating to be higher or lower than that due to chance error? [Source: Frank Newport, Lydia Saad, and David Moore, “How Polls Are Conducted,” in *Where America Stands* (New York: John Wiley & Sons, Inc., © 1997). Used by permission of John Wiley & Sons, Inc.]

E98. Some people complain that election polls cannot be right, because they personally were not asked how they voted. Write an explanation about how polls can get a good idea of how the entire population will vote by asking a relatively small number of voters.
AP1. Only 33% of students correctly answered a difficult multiple-choice question on an exam given nationwide. Ms. Chang gave the same question to her 35 students, hypothesizing that they would do better than students nationwide. Despite the lack of randomization, she performed a one-sided test of the significance of a sample proportion and got $P = 0.03$. Which is the best interpretation of this $P$-value?

A. Only 3% of her students scored better than students nationally.
B. If the null hypothesis is true that her students do the same as students nationally, there is a 3% chance that her students will do better than students nationally on this question.
C. Between 30% and 36% of her population of students can be expected to answer the question correctly.
D. There is a 3% chance that her students are better than students nationally.
E. There is a 3% chance that a random sample of 35 students nationwide would do as well as or better than her students did.

AP2. Researchers constructed a 95% confidence interval for the proportion of people who prefer apples to oranges. They computed a margin of error of $4\%$. In checking their work, they discovered that the sample size used in their computation was $\frac{1}{3}$ of the actual number of people surveyed. Which is closest to the correct margin of error?

A. 1%  
B. 2%  
C. 4%  
D. 8%  
E. 16%

AP3. A survey of 200 randomly selected alumni of Lincoln High School found that 105 favor abolishing the school dress code. Is this convincing evidence that more than half of all alumni favor abolishing the dress code?

A. Yes, because 105 is more than half of 200.
B. Yes, by constructing a 95% confidence interval, you can see that it is plausible that more than half of all alumni favor abolishing the dress code.
C. No, by constructing a 95% confidence interval, you can see that it is plausible that 50% or less of all alumni favor abolishing the dress code.
D. No, because the survey only used a sample of alumni.
E. Because alumni probably have strong opinions that aren't normally distributed, no conclusion can be reached.

AP4. In a study of the effectiveness of two tutors in preparing students for an exam, 50 students were randomly assigned either to Mr. A or to Mr. B. A larger proportion of students tutored by Mr. A passed the exam, resulting in a 95% confidence interval for the difference of two proportions of $(-0.05, 0.45)$. Which is a correct conclusion to draw from this?

A. There is statistically significant evidence that Mr. A is the better tutor.
B. You cannot reject a null hypothesis of equal tutor effectiveness.
C. The difference in the percentage who passed the exam and were tutored by Mr. A and the percentage who passed and were tutored by Mr. B is 20%.
D. A Type II error may be made.
E. The design of this study didn't have enough power to pick up any but a very large difference in passing rates.

AP5. With $\alpha = 0.05$, researchers conducted a test of the difference of two proportions to compare the rate of alcohol use among teens this year and in 1990. The rates for both years are based on large, independent random samples of teens. Which is the best interpretation of “$\alpha = 0.05$” in this context?

A. There’s a 5% chance that the rate of alcohol use has changed.
B. There’s a 5% chance that the rate of alcohol use has not changed.
C. If the rate of alcohol use has not changed, there’s a 5% chance that the researchers will mistakenly conclude that it has.
If the rate of alcohol use has changed, there’s a 5% chance that the researchers will mistakenly believe it hasn’t. The study has enough power to detect a difference in the rate of alcohol use of 0.05 or more.

AP6. A student obtains a random sample of M&M’s and a random sample of Skittles. She finds that 7 of the 40 M&M’s are yellow and 13 of the 35 Skittles are yellow. Her null hypothesis is that the proportion of yellow candies is equal in both brands. Her alternative hypothesis is that the proportion is higher in Skittles. What is her conclusion, if she uses $\alpha = 0.05$?

- Reject the null hypothesis. The difference in the two proportions can reasonably be attributed to chance alone.
- Reject the null hypothesis. The difference in the two proportions cannot reasonably be attributed to chance alone.
- Do not reject the null hypothesis. The difference in the two proportions can reasonably be attributed to chance alone.
- Do not reject the null hypothesis. The difference in the two proportions cannot reasonably be attributed to chance alone.
- Accept the null hypothesis because the difference in the two proportions is not statistically significant.

AP7. In a pre-election poll, 51% of a random sample of voters plan to vote for the incumbent. A 95% confidence interval was computed for the proportion of all voters who plan to vote for the incumbent. What is the best meaning of “95% confidence”?

- If all voters were asked, there is a 95% chance that 51% of them will say they plan to vote for the incumbent.
- You are 95% confident that the confidence interval contains 51%.
- In 100 similar polls, you expect that 95 of the confidence intervals constructed will contain the percentage of all voters who plan to vote for the incumbent.
- In 100 similar polls, you expect that 95 of the confidence intervals constructed will contain 51%, which is the best estimate of the percentage of all voters who plan to vote for the incumbent.
- The probability is 0.95 that the confidence interval constructed in this poll contains the percentage of all voters who plan to vote for the incumbent.

AP8. Sheldon takes a random sample of 50 U.S. housing units and finds that 30 are owner-occupied. Using a significance test for a proportion, he is not able to reject the null hypothesis that exactly half of U.S. housing units are owner-occupied. Later, Sheldon learns that the U.S. Census for the same year found that 66.2% of housing units are owner-occupied. Select the best description of the type of error in this situation.

- No error was made.
- A Type I error was made because a false null hypothesis was rejected.
- A Type I error was made because a false null hypothesis wasn’t rejected.
- A Type II error was made because a false null hypothesis was rejected.
- A Type II error was made because a false null hypothesis wasn’t rejected.

Investigative Tasks

AP9. Fifty students want to know what percentage of Skittles candies are orange. Each student takes a random sample of 50 candies and constructs a 90% confidence interval using the formula

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

a. Of all Skittles candies, 20% are orange. How many of their fifty 90% confidence intervals would you expect to capture the population proportion of 0.20?
b. The students’ confidence intervals are plotted in Display 8.36 as vertical lines. How many of their confidence intervals capture the population proportion of 0.20? What percentage is this?

c. Are most of the confidence intervals that don’t capture the population proportion too close to 0 or too close to 0.5? Do they tend to be shorter or longer than intervals that do capture the population proportion?

d. Unfortunately, what you have discovered here is true in general: If you use the standard formula to construct confidence intervals for a proportion, your actual capture rate will be smaller (and sometimes a lot smaller) than advertised. List several ways in which this formula is based on approximations and so might have an actual capture rate different from that advertised.

AP10. The Plus 4 interval. In AP9, you saw that the standard formula for a confidence interval results in a capture rate that tends to be less than advertised. The capture rate should improve if the more extreme intervals are moved toward the center and made a little longer, as long as the remaining intervals aren’t disturbed too much. There is a method that will do exactly that, called the Plus 4 method. You simply add 2 to the number of observed successes and add 2 to the number of observed failures (which adds 4 to the sample size). That is, if \( x \) denotes the number of successes in a sample of \( n \) items, the new estimator of the population proportion, \( \hat{p} \), is

\[
\hat{p} = \frac{x + 2}{n + 4}
\]

The Plus 4 confidence interval is then

\[
\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n + 4}}
\]

a. Construct 90% confidence intervals for \( n = 50 \) and \( x = 5 \) using both the standard formula and the Plus 4 method. What happened to the center of the interval? The length of the interval?

b. Repeat the procedures and questions of part a for \( n = 50 \) and \( x = 20 \). Compare your answers to those in part a.

c. The students in AP9 used the same samples as before to construct Plus 4 confidence intervals. These Plus 4 confidence intervals are given in Display 8.37. How many of the intervals capture the population proportion of 0.20? What percentage is this?

d. Describe how the Plus 4 confidence intervals differ from the standard intervals.
Inference for Means

Does it make a difference whether you take water samples at mid-depth or near the bottom when studying pesticide levels in a river? Inference for differences is fundamental to statistical applications because many surveys and all experiments are comparative in nature.

<table>
<thead>
<tr>
<th>Bottom</th>
<th>Mid-depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>28</td>
</tr>
<tr>
<td>3</td>
<td>389</td>
</tr>
<tr>
<td>4</td>
<td>367</td>
</tr>
<tr>
<td>5</td>
<td>22</td>
</tr>
<tr>
<td>6</td>
<td>236</td>
</tr>
<tr>
<td>7</td>
<td>33</td>
</tr>
<tr>
<td>8</td>
<td>81</td>
</tr>
<tr>
<td>9</td>
<td>68</td>
</tr>
</tbody>
</table>

Concentration of Aldrin
You now know how to make inferences for proportions based on categorical data. But often the data that arise in everyday contexts are measurement data rather than categorical data. Measurement data, as you know, are most often summarized by using the mean. Mean income, mean scores on exams, mean waiting times at checkout lines, and mean heights of people your age are all commonly used in making decisions that affect your life.

In Chapter 8, you learned how to construct confidence intervals and perform significance tests for proportions. Although the formulas you’ll use for the inferential procedures in this chapter will change a little from those you just learned, the basic concepts remain the same. These concepts are all built around the question “What are the reasonably likely outcomes from a random sample?” This chapter will follow the same outline as Chapter 8 except that the methods are applied to means rather than proportions.

In this chapter, you will learn how to

• construct and interpret a confidence interval for estimating an unknown mean
• perform a significance test for a single mean
• construct and interpret a confidence interval for estimating the difference between two means
• perform a significance test for the difference between two means
• use a confidence interval to estimate the mean of the differences from paired samples
• perform a significance test for the mean of the differences from paired samples
A Confidence Interval for a Mean

In this section you will learn how to construct a confidence interval for a population mean, $\mu$. The structure of the confidence interval will be similar to that in Section 8.1 for a proportion:

$$\hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = \text{statistic} \pm (\text{critical value}) \cdot (\text{standard deviation of statistic})$$

Let’s start with an example of a real-life situation in which it is important to locate the mean of a population.

**Example: Average Body Temperature**

When you are sick and take your temperature, you probably compare your temperature to the so-called “normal” temperature, 98.6°F. How was that temperature determined? In the 1860s, the German physician Carl Wunderlich measured the temperatures of thousands of healthy people and reported the mean as 37°C (98.6°F). Investigators have questioned his methodology and the quality of his thermometers. To determine an up-to-date average, researchers took the body temperatures of 148 people at several different times during two consecutive days. A portion of these data, for ten randomly selected women, is given here (in °F):

97.8 98.0 98.2 98.2 98.2 98.6 98.8 98.8 99.2 99.4


The mean body temperature, $\bar{x}$, for this sample of ten women is 98.52, and the standard deviation, $s$, is 0.527. Are these statistics likely to be equal to the mean $\mu$ and standard deviation $\sigma$ for the population? How can you determine the plausible values of the mean temperature of all women?

**Solution**

Neither $\bar{x}$ nor $s$ is likely to be exactly equal to the population parameters $\mu$ and $\sigma$. (This would be true no matter how large the sample.)

Plausible values of the mean body temperature of all women, $\mu$, are those values that lie “close” to $\bar{x} = 98.52$, where “close” is defined in terms of standard error. From Section 7.3, the standard error of the sampling distribution of a sample mean is given by $SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, where $\sigma$ is the standard deviation of the population and $n$ is the sample size. When the sample size is large enough or the population is normally distributed, in 95% of all samples $\bar{x}$ and $\mu$ are no farther apart than $1.96 \cdot \frac{\sigma}{\sqrt{n}}$. So plausible values of $\mu$ lie in the interval

$$\bar{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad 98.52 \pm 1.96 \cdot \frac{0.527}{\sqrt{10}}$$
For the sample of body temperatures, you know $\bar{x}$ and $n$, but you don’t know $\sigma$—you seldom will know $\sigma$ in real-life situations. What now? As you might be thinking, you will have to use the sample standard deviation, $s$, as an estimate of $\sigma$.

**The Effect of Estimating $\sigma$**

To compute a confidence interval for the mean, you use the sample standard deviation, $s$, as an estimate of $\sigma$. How will substituting $s$ for $\sigma$ affect your confidence interval? Some samples give an estimate that’s too small: $s < \sigma$. In this case, the confidence interval will be too narrow. Others give an estimate that’s too large: $s > \sigma$. In this case, the confidence interval will be too wide.

On average, the sampling distribution of $s$ has its center very near $\sigma$, so, on average, the width of the confidence interval is about right. That’s a nice feature, but confidence intervals are judged on their capture rate, not their average width. Does replacing $\sigma$ with $s$ change your chance of capturing the unknown population mean? You will investigate this matter in Activity 9.1a.

---

**ACTIVITY 9.1a**

The Effect of Estimating the Standard Deviation on Measuring Romantic Love

Nerve growth factor (NGF) is a neurotrophin found in blood plasma. Researchers found that the mean NGF concentration of people “truly, deeply and madly in love” for six months or less was significantly higher than that in single people or in people in a long-term relationship. The NGF levels in those who recently fell in love were normally distributed with mean 227 (pg/mL) and standard deviation 107. Assume that these are the population parameters.


Suppose you must estimate the population mean from the NGF levels of a random sample of four people who recently fell in love.

1. Use your calculator to generate a random sample of four NGF levels from this normally distributed population, and store them into list 1. [See Calculator Note 7A.]

```
randNorm(227, 107,
4) → L₁
```

2. Use your calculator to find the mean, $\bar{x}$, and standard deviation, $s$. Record the value of $s$. [See Calculator Note 2F.]

3. Construct a confidence interval using the formula

$$\bar{x} \pm 1.96 \cdot \frac{s}{\sqrt{n}}$$

Does your confidence interval capture the population mean, 227?

(continued)
4. Follow the procedure in steps 1 through 3 until the students in your class have generated a total of 100 samples. You will have 100 values of $s$ and 100 decisions as to whether the confidence interval captured the population mean, $\mu$.

5. Make a dot plot of the 100 values of $s$ and describe its shape, mean, and spread. How close to $\sigma$ is the mean of the values of $s$? How many times is $s$ smaller than $\sigma$? Larger?

6. What percentage of your confidence intervals should capture the true population mean, $\mu$, of 227? What percentage of your confidence intervals actually capture $\mu$? What can you conclude about the effect of using $s$ to estimate $\sigma$?

You learned in the activity that the sampling distribution of $s$ is skewed right. Although the average value of $s$ is about equal to $\sigma$, $s$ tends to be smaller than $\sigma$ more often than it is larger because the median is smaller than the mean for this skewed distribution. This causes the confidence intervals to be too narrow more than half the time, so the capture rate will be less than the advertised value of 95%.

Fortunately, the sampling distribution of $s$ becomes less skewed and more approximately normal as the sample size increases. The histogram on the left in Display 9.1 shows the values of $s$ for 1000 samples of size 4 taken from a normally distributed population with $\sigma = 107$. For this small sample size, the distribution is quite skewed. The histogram on the right shows the values of $s$ for 1000 samples of size 20. There is little skewness here, so $s$ is smaller than $\sigma$ about as often as it is larger.

![Histograms showing sampling distribution of $s$](image)

**Display 9.1** Approximate sampling distribution of $s$ for samples of size 4 (left) and size 20 (right) for a normally distributed population with $\sigma = 107$.

**DISCUSSION**

### The Effect of Estimating $\sigma$

D1. Consider two extreme situations, $n = 10$ and $n = 1000$. If you use $s$ for $\sigma$ in the formula, $\bar{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}$, which sample size—10 or 1000—do you expect to give a capture rate closest to 95%? Why?
D2. When you use $s$ to estimate $\sigma$, the capture rate is too small unless you make further adjustments. If an interval's capture rate is smaller than what you want it to be, do you need to use a wider or a narrower interval to get the capture rate you want?

### How to Adjust for Estimating $\sigma$

As you have seen, using $s$ as an estimate of $\sigma$ in a confidence interval lowers the overall capture rate because $s$ tends to be smaller than $\sigma$ more often than it is larger. To compensate, you will increase the width of the confidence interval by replacing $z^*$ with a larger value, called $t^*$:

$$\bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$$

Let's listen in on a discussion about $t^*$.

**Student:** Where does the value of $t^*$ come from?

**Statistician:** In principle, you could find it using simulation. Set up an approximately normal population, take a random sample, and compute the mean and standard deviation. Do this thousands of times. Then use the results to figure out the value of $t^*$ that gives a 95% capture rate for intervals of the form $\bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$.

**Student:** Wouldn't that take a lot of work?

**Statistician:** Yes, especially if you went about it by hand. Fortunately, this work has already been done, long ago. Statistician W. S. Gosset (1876–1937), who worked for the Guinness Brewery, actually did this back in 1915. Four years later, geneticist and statistician R. A. Fisher (1890–1962) figured out how to find values of $t^*$ using calculus. If the population is nearly normal, it turns out that the value of $t^*$ depends on just two things: how many observations you have, and the capture rate you want.

**Student:** So $t^*$ doesn’t depend on the unknown mean or unknown standard deviation?

**Statistician:** No, it doesn’t, which is very handy because in practice you don’t know these numbers. Suppose, for example, you have a sample of size 5 and you want a 95% confidence interval. Then you can use $t^* = 2.776$ no matter what the values of $\mu$ and $\sigma$ are.

**Student:** Where did you get that value for $t^*$?

**Statistician:** From a $t$-table, although I could have found it from a calculator or statistical software. A brief version of the table is shown in Display 9.2. Table B on page 826 is more complete. For 95% confidence, you go to the column with 95% at the bottom. For the row, you need to know the degrees of freedom, $df$ for short.
Tail Probability $p$

<table>
<thead>
<tr>
<th>df</th>
<th>.05</th>
<th>.025</th>
<th>.02</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.314</td>
<td>12.71</td>
<td>15.89</td>
</tr>
<tr>
<td>2</td>
<td>2.920</td>
<td>4.303</td>
<td>4.849</td>
</tr>
<tr>
<td>3</td>
<td>2.353</td>
<td>3.182</td>
<td>3.482</td>
</tr>
<tr>
<td>4</td>
<td>2.132</td>
<td>2.776</td>
<td>2.999</td>
</tr>
<tr>
<td>5</td>
<td>2.015</td>
<td>2.571</td>
<td>2.757</td>
</tr>
</tbody>
</table>

Confidence level $C$

Display 9.2. An abbreviated version of a t-table, with a sketch showing the shaded tail areas.

**Student:** Degrees of freedom? What's that?

**Statistician:** There's a short answer, a longer answer, and a very long answer. The longer answer will come in E40. The very long answer is for another course. For the moment, here's the short answer: The number of **degrees of freedom** is one less than the sample size. (It is the number you use for the denominator when you calculate the standard deviation, $s$.) So for these confidence intervals, $df = n - 1$, where $n$ is the sample size. When $n = 5$, $df = 4$, and you look in that row of the table. If you look in the row with $df = 4$ and the column with confidence level 95%, you'll find the value 2.776 for $t^*$.

**Finding the Value of $t^*$ to Use in Your Confidence Interval**

Turn to Table B on page 826. Find your confidence level in the last row of the table. Find your $df$, $n - 1$ where $n$ is the sample size, in the column at the left. Your value of $t^*$ is at the intersection of this row and column. If your $df$ falls between two of those listed in the column on the left, use the smaller $df$.

[To find $t^*$ using a calculator, see **Calculator Note 9A**.]

**Example: Who Is Hotter?**

Is normal body temperature the same for men and women? As you saw in the previous example, medical researchers interested in this question collected data from a large number of men and women. Random samples from that data are recorded in Display 9.3. The women's temperatures are the same as those given in the previous example.

a. Use a 95% confidence interval to estimate the mean body temperature of men.

b. Use a 95% confidence interval to estimate the mean body temperature of women.
Solution

First, you need to check that these are random samples from relatively large populations and that it’s reasonable to assume that the populations are normal. (More on this is in Section 9.3.) The problem states that these are random samples selected from a group of men and women examined by the researchers, so your final result generalizes only to this population. The researchers reported that the distributions are normal. You should always plot the data, however, and the stemplots in Display 9.4 give you no reason to suspect that the populations are not normally distributed.

For a sample of size 10 (9 degrees of freedom) and a confidence level of 95%, Table B on page 826 gives a value of \( t^* \) of 2.262.

a. For the males, the sample mean is 97.88, and the sample standard deviation is 0.555. That gives a 95% confidence interval of

\[
\bar{x}_m \pm t^* \cdot \frac{s_m}{\sqrt{n}} \quad \text{or} \quad 97.88 \pm 2.262 \cdot \frac{0.555}{\sqrt{10}}
\]

or (97.48, 98.28). Any population of body temperatures with a mean in this interval could have produced a sample mean of 97.88 as a reasonably likely outcome.
b. For the females, the sample mean is 98.52, and the sample standard deviation is 0.527. You are 95% confident that the mean body temperature of women in this population is in the interval

\[ \bar{x}_f \pm t^* \frac{s_f}{\sqrt{n}} \quad \text{or} \quad 98.52 \pm 2.262 \frac{0.527}{\sqrt{10}} \]

or (98.14, 98.90).

From the sample means it might appear that the average body temperature of females is higher than that of males, but beware! There is some overlap in the confidence intervals, so that conclusion might not be valid. You’ll come back to the question of comparing two means in Section 9.4.

**DISCUSSION**

**How to Adjust for Estimating \( \sigma \)**

D3. What parameter does the standard deviation, \( s \), of a random sample tend to approach as the sample size gets larger? Look at Table B on page 826. What happens to the value of \( t^* \) as the sample size increases? Find the row where \( df \) equals \( \infty \). Do those numbers look familiar?

D4. Overall, the researchers found a mean body temperature of about 98.2°F, which is lower than Wunderlich’s 98.6°F that has become the standard. Was Wunderlich wrong? Wunderlich worked in the metric system. He rounded his mean body temperature to the nearest degree to get 37°C. What is the lowest his mean could have been in degrees Celsius? The highest? Convert the lowest possible mean, the reported mean (37°C), and the highest possible mean to degrees Fahrenheit. The conversion formula is \( F = \frac{9}{5} C + 32 \). What can you conclude?

**Constructing a Confidence Interval for a Mean**

Whenever you construct a confidence interval based on \( t \), there are three conditions you must check. Officially, you need

- a random sample (or random assignment of treatments to units)
- an approximately normally distributed population or a large enough sample size that the distribution of the sample mean is approximately normal
- in the case of a survey, the size of the population should be at least ten times the size of the sample

The fact that you must always have a random sample should be no surprise. A normally distributed population is (theoretically) required because the values of \( t^* \) in Table B were computed by assuming that the population is normally distributed. However, you can get away with less than a normally distributed population if the sample size is large enough (this will be explained in Section 9.3). If the sample size is more than about 10% of the population size, the estimated standard error of the mean used in this chapter will tend to be larger than it needs to be for the specified confidence level.
That's the last of the new things to learn about constructing a confidence interval for a mean. Here's a summary of what you've learned.

A Confidence Interval for a Mean

Check conditions.
You must check three conditions:

- Randomness. In the case of a survey, the sample must have been randomly selected. In the case of an experiment, the treatments must have been randomly assigned to the experimental units.

- Normality. The sample must look like it's reasonable to assume that it came from a normally distributed population or the sample must be large enough that the sampling distribution of the sample mean is approximately normal. There is no exact rule for determining whether a normally distributed population is a reasonable assumption or what constitutes a “large enough” sample, but you will learn some guidelines in Section 9.3.

- Sample Size. In the case of a sample survey, the population size should be at least ten times as large as the sample size.

Do computations.
A confidence interval for the population mean, \( \mu \), is given by

\[
\bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}
\]

where \( n \) is the sample size, \( \bar{x} \) is the sample mean, \( s \) is the sample standard deviation, and \( t^* \) depends on the confidence level desired and the degrees of freedom, \( df = n - 1 \).

Give interpretation in context.
A good interpretation is of this form: “I am 95\% confident that the population mean, \( \mu \), is in this confidence interval.”

Of course, when you interpret a confidence interval, you will do it in context, describing the population you are talking about.

Example: The Atomic Weight of Silver

The National Institute of Standards and Technology (NIST) conducts research that helps U.S. industry measure, test, and standardize products and services. In order to reduce uncertainty in the atomic weight of silver, an NIST researcher measured the atomic weight of a reference sample of silver. The 24 measurements from one mass spectrometer are given in Display 9.5. Note that all 24 measurements begin with 107.868; thus, it will be easier to work with the coded values, the last four decimal places.
Construct and interpret a 95% confidence interval for the mean measurement (using this mass spectrometer) of the atomic weight of silver.

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Coded Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>107.8681079</td>
<td>1079</td>
</tr>
<tr>
<td>107.8681344</td>
<td>1344</td>
</tr>
<tr>
<td>107.8681513</td>
<td>1513</td>
</tr>
<tr>
<td>107.8681197</td>
<td>1197</td>
</tr>
<tr>
<td>107.8681604</td>
<td>1604</td>
</tr>
<tr>
<td>107.8681385</td>
<td>1385</td>
</tr>
<tr>
<td>107.8681642</td>
<td>1642</td>
</tr>
<tr>
<td>107.8681365</td>
<td>1365</td>
</tr>
<tr>
<td>107.8681151</td>
<td>1151</td>
</tr>
<tr>
<td>107.8681082</td>
<td>1082</td>
</tr>
<tr>
<td>107.8681517</td>
<td>1517</td>
</tr>
<tr>
<td>107.8681448</td>
<td>1448</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Coded Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>107.8681198</td>
<td>1198</td>
</tr>
<tr>
<td>107.8681482</td>
<td>1482</td>
</tr>
<tr>
<td>107.8681334</td>
<td>1334</td>
</tr>
<tr>
<td>107.8681609</td>
<td>1609</td>
</tr>
<tr>
<td>107.8681101</td>
<td>1101</td>
</tr>
<tr>
<td>107.8681512</td>
<td>1512</td>
</tr>
<tr>
<td>107.8681469</td>
<td>1469</td>
</tr>
<tr>
<td>107.8681360</td>
<td>1360</td>
</tr>
<tr>
<td>107.8681254</td>
<td>1254</td>
</tr>
<tr>
<td>107.8681261</td>
<td>1261</td>
</tr>
<tr>
<td>107.8681450</td>
<td>1450</td>
</tr>
<tr>
<td>107.8681368</td>
<td>1368</td>
</tr>
</tbody>
</table>

Display 9.5 Twenty-four measurements of a sample of silver.


Solution

Check conditions.

You can consider these measurements to be a random sample of all possible measurements by this mass spectrometer. Display 9.6 shows that the distribution of measurements from the sample is reasonably symmetric with no outliers, so it is reasonable to assume that they could have come from a normal distribution.

Display 9.6 A dot plot of the sample data.

Do computations.

The mean and standard deviation of the measurements are 1363.54 and 169.017, respectively. For 95% confidence and \( df = 24 - 1 = 23 \), use \( t^* = 2.069 \). The 95% confidence interval is

\[
\bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}} = 1363.54 \pm 2.069 \cdot \frac{169.017}{\sqrt{24}} 
\]

\[
\approx 1363.54 \pm 71.38
\]

or about (1292, 1435).

You are 95% confident that the mean measurement of the atomic weight of silver using this mass spectrometer is in the interval from 107.8681292 to 107.8681435.
You also can use your calculator to calculate the confidence interval for a mean, given either the original data or the summary statistics of the mean and standard deviation [See Calculator Note 9B]. Both approaches are shown here for the data in the previous example.

### Constructing a Confidence Interval for a Mean

D5. A student said, “Let’s be sure and capture \( \mu \) by producing an interval with a 100% confidence level.” Is this possible? Would it be wise to shoot for a 99.5% confidence level?

D6. You have produced a confidence interval for your supervisor, and she says, “This interval is too wide to be of any practical value.” What are your options for producing a narrower interval in the next study of this same type? Which option would you choose if the study might have consequences that are vital to the future of the firm?

### More on Interpreting a Confidence Interval

As in Chapter 8, the capture rate of a confidence interval is the proportion of random samples for which the resulting confidence interval captures the (population) parameter. You can think of it as the chance that the method used to produce a confidence interval actually will work correctly.

### The Capture Rate of a Confidence Interval for a Mean

The proportion of intervals of the form \( \bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}} \) that capture \( \mu \) is equal to the confidence level. That is, a 95% confidence interval will have a 95% capture rate, a 90% confidence interval will have a 90% capture rate, and so on.

This will be true provided that

- the sample is random
- the population is normally distributed
- the size of the population is at least 10 times the sample size

The capture rate will be approximately correct for non-normally distributed populations as long as the sample size is large enough.
Remember that the capture rate is a property of the method, not of the sample. After the sample is selected, either the resulting interval will capture the population mean or it won't. So it is incorrect to say about a specific interval that there is a 95% probability (or 95% chance) that the true mean falls in this interval. All you can say is that you are 95% confident that \( \mu \) is in your interval, meaning that the method produces a confidence interval that captures \( \mu \), the true mean, 95% of the time.

### The Meaning of “95% Confident”

Saying you are “95% confident” means that if you could take random samples repeatedly from this population and compute a confidence interval from each sample, in the long run 95% of these different intervals would contain (or capture) \( \mu \).

In other words, a 95% confidence interval for a population mean, \( \mu \), includes the plausible values of \( \mu \). A population with any one of these values for \( \mu \) could have produced the observed sample mean as a reasonably likely outcome.

In practice, samples might not be random, distributions might not be normal, and survey questions might be worded so that people answer incorrectly. In such situations, the capture rate may be far below what is advertised.

### More on Interpreting a Confidence Interval

D7. Explain why it is incorrect to say you are 95% confident that the population mean is in the confidence interval that you just constructed.

D8. “The capture rate equals the confidence level.” Explain why this statement depends on

a. the randomness of the sampling

b. the normality of the population

### Margin of Error

The quantity

\[
E = t^* \cdot \frac{s}{\sqrt{n}}
\]

is called the margin of error. It is half the width of the confidence interval. When your samples are random, larger samples provide more information than do smaller ones. So, in general, as the sample size gets larger, the margin of error gets smaller.

#### Example: Margin of Error for the Mean Atomic Weight of Silver

Find the margin of error for the mean atomic weight of silver in the example on pages 569–570.
Solution

The confidence interval for the mean atomic weight was

\[ \bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}} = 1363.54 \pm 2.069 \cdot \frac{169.017}{\sqrt{24}} \]

\[ \approx 1363.54 \pm 71.38 \]

So the margin of error is 71.38.

Summary 9.1: A Confidence Interval for a Mean

In this section you have learned how to construct a confidence interval for a population mean, \( \mu \). A confidence interval for \( \mu \) is given by

\[ \bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}} \]

where \( \bar{x} \) is the sample mean, \( s \) is the sample standard deviation, and \( n \) is the sample size. Because you are using \( s \) to estimate \( \sigma \), you must use values of \( t^* \) rather than values of \( z^* \). Find the value of \( t^* \) in Table B on page 826 or on your calculator, using \( df = n - 1 \). For a given confidence level and sample size, the value of \( t^* \) is slightly larger than the corresponding value of \( z^* \), but this difference becomes smaller with larger sample sizes. You must check that you have a random sample from a relatively large population (or a random assignment of treatments to units) and that the sample came from a distribution that is approximately normal or, if not, that the sample is large enough.

Here is how to interpret this confidence interval:

- You are 95% confident that the population mean, \( \mu \), is in this confidence interval.
- This method of constructing a 95% confidence interval ensures that the chance of getting a value of \( \bar{x} \) whose interval captures \( \mu \) is 95%. This is called the capture rate or confidence level.

Practice

The Effect of Estimating \( \sigma \)

P1. *Aldrin in the Wolf River.* Aldrin is a highly toxic organic compound that can cause various cancers and birth defects. Ten samples taken from Tennessee’s Wolf River downstream from a toxic waste site once used by the pesticide industry gave these aldrin concentrations, in nanograms per liter:

<table>
<thead>
<tr>
<th>Concentration (ng/L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.17</td>
</tr>
<tr>
<td>3.76</td>
</tr>
</tbody>
</table>

a. Calculate a confidence interval of the form \( \bar{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}} \), using \( s \) as an estimate of \( \sigma \).

b. In general, how do you expect this estimate—using \( s \) for \( \sigma \)—to affect the center of a confidence interval? Will it tend to make the interval too wide or too narrow? Will it make the capture rate larger than 95% or smaller than 95%?

How to Adjust for Estimating \( \sigma \)

P2. *Using a t-Table.* Use Table B on page 826 to find the correct value of \( t^* \) to use for each of these situations.

a. A 95% confidence interval based on a sample of size 10
b. a 95% confidence interval based on a sample of size 7

c. a 99% confidence interval based on a sample of size 12

d. a 90% confidence interval with \( n = 3 \)

e. a 99% confidence interval with \( n = 44 \)

f. a 99% confidence interval with \( n = 82 \)

P3. For each situation, construct a 95% confidence interval for the unknown mean.

a. \( n = 4 \) \( \bar{x} = 27 \) \( s = 12 \)

b. \( n = 9 \) \( \bar{x} = 6 \) \( s = 3 \)

c. \( n = 16 \) \( \bar{x} = 9 \) \( s = 48 \)

P4. Using the data in P1, construct a 95% confidence interval for the mean level of aldrin. Compare the width of this confidence interval to your interval in P1.

Constructing a Confidence Interval for a Mean

P5. An article in the *Journal of the American Medical Association* included the body temperatures of 122 men. (The data in the example on pages 566–567 were a random sample taken from this larger group of men.) The summary statistics for the entire sample of men's temperatures were \( \bar{x} = 98.1, \) \( s = 0.73, \) and \( n = 122. \) [Source: P.A. Mackowiak et al., "A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich," *Journal of the American Medical Association* 268 (September 1992): 1578–80.]

a. Construct a 95% confidence interval for the mean of the population from which these data were selected. You can assume that the conditions have been met.

b. Write an interpretation of this confidence interval.

c. Is 98.6°F in this interval? What can you conclude?

P6. The data in Display 9.7 are from a survey taken in an introductory statistics class during the first week of the semester. These data are the number of hours of study per week reported by each of the 61 students. Because of the way students register for class, these students can be considered a random sample of all students taking this course.

<table>
<thead>
<tr>
<th>Study Hours</th>
<th>n</th>
<th>MEAN</th>
<th>MEDIAN</th>
<th>STDEV</th>
</tr>
</thead>
<tbody>
<tr>
<td>61</td>
<td>10.26</td>
<td>10</td>
<td>6.22</td>
<td></td>
</tr>
</tbody>
</table>

Display 9.7 Number of hours of study per week.

a. You were told that the sample can be considered a random selection from this large population of students. Do the data look as if they reasonably could have come from a normal distribution?

b. Regardless of your answer to part a, estimate the mean study hours per week for all the students taking this course, with 90% confidence.

c. Explain the meaning of your confidence interval.

More on Interpreting a Confidence Interval

P7. Refer to the 95% confidence interval you constructed for the aldrin data in P4. Which of statements A–E are correct interpretations of that interval?

A. If you take one more measurement from the Wolf River, you are 95% confident that this measurement will fall in the confidence interval.

B. If you take ten more measurements from the Wolf River, you are 95% confident that the sample mean of the ten measurements will fall in the confidence interval.

C. There is a 95% chance that the mean aldrin level of the Wolf River falls in the confidence interval.
D. You are 95% confident that the mean aldrin level of the Wolf River falls in the confidence interval.

E. You are 95% confident that the sample mean falls in the confidence interval.

P8. A 95% confidence interval for a population mean is calculated for a random sample of weights, and the resulting confidence interval is from 42 to 48 lb. For each statement, indicate whether it is a true or false interpretation of the confidence interval. Explain why the false statements are false.

a. 95% of the weights in the population are between 42 lb and 48 lb.

b. 95% of the weights in the sample are between 42 lb and 48 lb.

c. The probability that the interval includes the population mean, μ, is 95%.

d. The sample mean, \( \bar{x} \), might not be in the confidence interval.

e. If 200 confidence intervals were generated using the same process, about 10 of these confidence intervals would not include the population mean.

Margin of Error

P9. Refer to the confidence intervals for the mean body temperature of men and women in the example on pages 566–567.

a. Find the margin of error for the men and for the women.

b. Which margin of error is larger? Why is it larger?

Exercises

E1. If you are lost in the woods and do not have clear directional markers, will you tend to walk in a circle? To study this question, some students recruited 30 volunteers to attempt to walk the length of a football field while blindfolded. Each volunteer began at the middle of one goal line and was asked to walk to the opposite goal line, a distance of 100 yards. None of them made it. Display 9.8 shows the number of yards they walked before they crossed a sideline.

a. Are the conditions met for finding a confidence interval estimate of the mean distance walked before crossing a sideline? Explain any reservations you have.

b. Regardless of your answer to part a, construct a 99% confidence interval for the mean distance walked before crossing a sideline.

c. Interpret your interval in part b.

<table>
<thead>
<tr>
<th>Yards Walked Before Walker Went Out of Bounds</th>
<th>Yards Walked Before Walker Went Out of Bounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>59</td>
</tr>
<tr>
<td>37</td>
<td>60</td>
</tr>
<tr>
<td>37</td>
<td>60</td>
</tr>
<tr>
<td>38</td>
<td>61</td>
</tr>
<tr>
<td>40</td>
<td>65</td>
</tr>
<tr>
<td>42</td>
<td>68</td>
</tr>
<tr>
<td>42</td>
<td>70</td>
</tr>
<tr>
<td>48</td>
<td>70</td>
</tr>
<tr>
<td>49</td>
<td>71</td>
</tr>
<tr>
<td>50</td>
<td>73</td>
</tr>
<tr>
<td>52</td>
<td>74</td>
</tr>
<tr>
<td>52</td>
<td>75</td>
</tr>
<tr>
<td>55</td>
<td>80</td>
</tr>
<tr>
<td>56</td>
<td>81</td>
</tr>
<tr>
<td>56</td>
<td>95</td>
</tr>
</tbody>
</table>

Display 9.8 Number of yards walked by blindfolded volunteers before going out of bounds.

E2. *Walking Babies.* Some babies start walking well before they are a year old, while others still haven’t taken their first unassisted steps at 18 months or even later. The data in Display 9.9 are from an experiment designed to see whether a program of special exercises for 12 minutes each day could speed up the process of learning to walk. In all, 23 week-old male infants (and their parents) took part in the study and were randomly assigned either to the special exercise group or to one of three control groups.

In order to isolate the effects of interest, the scientists used three different control groups. In the “exercise-control” group, parents were told to make sure their infant sons exercised at least 12 minutes per day but were given no special exercises and no other instructions about exercise. In the “weekly report” group, parents were given no instructions about exercise, but each week they were called to find out about their baby’s progress. Parents in the “final report” group also were given no instructions about exercise, nor did they receive weekly check-in calls. Instead, they gave a report at the end of the study. The response is the age, in months, when the baby first walked without help.

<table>
<thead>
<tr>
<th>Group</th>
<th>Age (in months) of First Unaided Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Special exercises</td>
<td>9, 9.5, 9.75, 10, 13, 9.5</td>
</tr>
<tr>
<td>Exercise-control</td>
<td>11, 10, 10, 11.75, 10.5, 15</td>
</tr>
<tr>
<td>Weekly report control</td>
<td>11, 12, 9, 11.5, 13.25, 13</td>
</tr>
<tr>
<td>Final report control</td>
<td>13.25, 11.5, 12, 13.5, 11.5</td>
</tr>
</tbody>
</table>


a. You can assume that there is no problem with the normality condition for constructing a confidence interval. Has the other condition been met?

b. Find a 95% confidence interval estimate of the mean time to first unaided steps for each treatment, assuming all 23 infants could have received that treatment. The 23 infants participating in the study constitute the population for this experiment.

c. Is it clear from your confidence intervals which treatment is best if you want baby boys to walk earlier on average?

E3. To avoid any further hill climbing, Jack and Jill have opened a water-bottling factory. The distribution of the number of ounces of water in the bottles is approximately normal. The mean, \( \mu \), is supposed to be 16 oz, but the water-filling machine slips away from that amount occasionally and has to be readjusted. Jack and Jill take a random sample of ten bottles from today’s production and weigh the water in each. The weights (in ounces) are

15.91  16.08  16.08  15.94  16.02  15.94  15.96  16.03  15.82  15.96

Should Jack and Jill readjust the machine? Use a statistical argument to support your advice.

E4. Display 9.10 gives the hot dog prices for three randomly selected National Basketball Association (NBA) teams.

<table>
<thead>
<tr>
<th>Team</th>
<th>Hot Dog Price (dollars)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boston</td>
<td>4.00</td>
</tr>
<tr>
<td>Atlanta</td>
<td>3.75</td>
</tr>
<tr>
<td>Dallas</td>
<td>3.75</td>
</tr>
</tbody>
</table>

Display 9.10  A sample of hot dog prices. [Source: www.teammarketing.com, March 2006.]

a. From this very small sample, is there any reason to suspect that the conditions are not met for constructing a confidence interval for the mean hot dog price for all NBA teams?

b. Regardless of your answer in part a, construct a 95% confidence interval.

c. The mean hot dog price for all NBA teams is $3.20. Is that value in your
A confidence interval? If not, list two reasons why this could have happened. The plots of all 30 prices in Display 9.11 might help you.

Display 9.11 Dot plot and boxplot of NBA hot dog prices.

d. There are only 30 teams in the NBA. If ten teams had been sampled instead of three, would the inferential methods of this section still have worked satisfactorily? If not, where do the potential problems lie?

E5. A statistics class at the University of Wisconsin–Stevens Point decided to estimate the average mass of a small bag of french fries sold at McDonald’s. They bought 32 bags during two different time periods on two consecutive days at the same McDonald’s and weighed the fries. The data are given in Display 9.12.

a. Are the conditions met for constructing a 95% confidence interval for the mean?

b. Regardless of your answer in part a, construct a 95% confidence interval.

c. The “target value” set by McDonald’s for the mass of a small order of fries was 74 g. Is there evidence that this McDonald’s franchise wasn’t meeting that target?

Display 9.12 Mass of small bags of McDonald’s fries. [Source: Nathan Wetzel, “McDonald’s French Fries. Would You Like Small or Large Fries?” STATS 43 (Spring 2005): 12–14.]

E6. The statistics class from E5 decided to estimate the average number of fries in a small bag of McDonald’s french fries. They bought 32 bags during two different time periods on two consecutive days at the same McDonald’s and counted the fries. The data are given in Display 9.13 (on the next page).

a. Are the conditions met for constructing a 95% confidence interval for the mean?

b. Regardless of your answer in part a, construct and interpret a 95% confidence interval.
Display 9.13  Number of fries in small bags of McDonald’s fries.

E7. A page was randomly selected from each of 30 brochures for cancer patients published by the American Cancer Society and the National Cancer Institute. The pages were judged for readability, using standard readability tests. The reading grade levels of the 30 pages selected are given in Display 9.14.

a. Suppose you want to estimate the mean reading level of cancer brochures. Do you have the required random sample?

b. Using the data in Display 9.14, construct a 90% confidence interval.

c. What is the best interpretation of your interval?


E8. Your grade in your literature class isn’t all that you had hoped it would be. You took five exams, each consisting of one essay question worth 100 points. Your scores were 52, 63, 72, 41, and 73. Your instructor averages these scores, gets 60.2, and says you have earned a D. You are, however, doing very well in your statistics class and think that what you have learned in this section can convince your literature instructor to give you a C, which is a mean score of between 70 and 79. You remember that your instructor told the class early in the semester that the questions on the exams would be only a sample of the questions that she could ask. Write a paragraph to your teacher giving your argument for a C. You can assume that the instructor knows something about statistics.

E9. Which confidence interval is widest in each of parts a–c? (You don’t need to do any computing.)

a. 95% interval with \( n = 4 \) and \( s = 10 \); 99% interval with \( n = 4 \) and \( s = 10 \)

b. 90% interval with \( n = 5 \) and \( s = 9 \)
b. 95% interval with \( n = 3 \) and \( s = 10 \);
95% interval with \( n = 4 \) and \( s = 10 \);
95% interval with \( n = 5 \) and \( s = 10 \)

c. 90% interval with \( n = 10 \) and \( s = 5 \);
95% interval with \( n = 10 \) and \( s = 5 \);
95% interval with \( n = 10 \) and \( s = 10 \)

E10. A large population of measurements has unknown mean and standard deviation. Two different random samples of 50 measurements are taken from the population. A 95% confidence interval for the population mean is constructed for each sample. Which statement would most likely be true of these two confidence intervals?

A. They would have identical values for the lower and upper limits of the confidence interval.
B. They would have the same margin of error.
C. The confidence intervals would have the same center but different widths.
D. None of the above is true.

E11. Suppose a large sample, with \( n = 100 \), is going to be taken from a population of 6-year-old girls. A 90% confidence interval will be constructed to estimate the population mean height. A smaller sample, with \( n = 50 \), will also be taken from the same population of 6-year-old girls, and a 99% confidence interval will be constructed to estimate the population mean height. Which of these statements is true about the chance of producing a confidence interval that captures the population mean height? Explain your reasoning.

A. The 90% confidence interval based on a sample of 100 heights has a better chance.
B. The 99% confidence interval based on a sample of 50 heights has a better chance.
C. Both methods have an equal chance.
D. You can't determine which method will have a better chance.

E12. Is the capture rate affected by

a. changes in the sample size?
b. the size of the population standard deviation?
c. the confidence level?

E13. What happens to the margin of error if all remains the same except that

a. the standard deviation of the sample is increased?
b. the sample size is increased?
c. the confidence level is increased?

E14. Refer to the four confidence intervals you constructed in E2.

a. Which confidence interval has the smallest margin of error?
b. Why is this the case?

E15. When you do not know the value of \( \sigma \) and have to estimate it with \( s \), the capture rate will be too small if you use \( z \) in your confidence interval. Complete this exercise to see part of the reason why.

a. In Activity 9.1a, you generated sample standard deviations, \( s \), from samples of size 4 taken from a population with known mean and standard deviation. Display 9.15 (on the next page) displays and summarizes 100 sample standard deviations, \( s \), for random samples of size 4 taken from a normal distribution with mean, \( \mu \), 511 and standard deviation, \( \sigma \), 112. What is the shape of this distribution? On average, how well does \( s \) approximate \( \sigma \)?

b. Which happened more often, a value of \( s \) smaller than \( \sigma \) or a value of \( s \) larger than \( \sigma \)?

c. How does your answer to part b explain why the capture rate tends to be too small if you use \( z \) in your confidence interval?
580 Chapter 9 Inference for Means

Step 1: If the population is normally distributed or the sample size is large enough, 95% of all sample means fall within 1.96 standard errors of the population mean.

Step 2: Writing this algebraically, for 95% of all sample means,

$$\mu - 1.96 \cdot SE < \bar{x} < \mu + 1.96 \cdot SE$$

(This interval contains the values of $\bar{x}$ that are reasonably likely.)

Step 3: You can write this in another, algebraically equivalent, way. For 95% of all sample means,

$$\bar{x} - 1.96 \cdot SE < \mu < \bar{x} + 1.96 \cdot SE$$

(This interval contains all the values of $\mu$ that are plausible as the population mean. In other words, each population mean in the interval $\bar{x} \pm 1.96 \cdot SE$ plausibly could produce the observed value of $\bar{x}$.)

Step 4: The inequality in step 3 can be rewritten using the fact that the standard error of the sampling distribution of the sample mean, $SE$, is $\sigma/\sqrt{n}$.

a. Show algebraically why the inequality in step 3 is equivalent to that in step 2.

b. Rewrite the inequality based on the information in step 4.

c. Explain why the inequality in part b is equivalent to a 95% confidence interval for the population mean.

### 9.2 A Significance Test for a Mean

Significance tests all rely on the same basic idea: Be suspicious of any model that assigns low probability to what actually happened. In this section you will learn when your data tell you to be suspicious of any claimed value of the population mean.

The basic procedure for testing a hypothesis is always the same: Compute a test statistic from the data and compare it with a known distribution. As
in Sections 8.2 and 8.4, you will check conditions for the test, state the null hypothesis and alternative hypothesis (two-sided or one-sided), find the $P$-value to assess the strength of the evidence against the null hypothesis, decide whether to reject the null hypothesis, and write a conclusion in context. You'll still have the possibility of making a Type I or Type II error, and sometimes you'll do a fixed-level test with a given significance level $\alpha$.

Activity 9.2a will help you get into the spirit of making an inference about a mean.

### ACTIVITY 9.2a

**What Makes Strong Evidence?**

The four data sets plotted in Display 9.16 each come from an experiment that measured the differences in pulse rates for pairs of people standing and sitting. In each experiment, two people with similar pulse rates were paired. One person was randomly assigned to stand and the other to sit. Then the difference in the two pulse rates, $standing - sitting$, was computed.

![Display 9.16 Differences in pulse rates, standing — sitting.](image)

1. Which experiment gives the strongest evidence in favor of a larger mean pulse rate when people are standing than when they are sitting?
2. Rank the four experiments from strongest to weakest in terms of the evidence provided. Explain your reasoning.
3. What else would you like to know about these data, in addition to what you see in the boxplots, in order to improve your inference-making ability?

### The Test Statistic

In Section 8.2, you computed this test statistic to test whether a hypothesized population proportion $p_0$ is consistent with the sample proportion, $\hat{p}$:

$$z = \frac{\text{statistic} - \text{parameter}}{\text{standard deviation of statistic}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$
The size of $z$ tells you how many standard errors your sample proportion $\hat{p}$ lies from the hypothesized proportion, $p_0$. Note that the standard error, $\sqrt{\frac{p_0(1-p_0)}{n}}$, can be computed because $p_0$ is stated in the null hypothesis ($H_0: p = p_0$) and $n$ is the known sample size.

What will change if you are testing a mean, $H_0: \mu = \mu_0$? It would be nice if you could use the $z$-statistic of Section 7.2 to measure how far a sample mean is from a hypothesized mean $\mu_0$:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

The catch is that you don't know $\sigma$, the standard deviation of the population.

Perhaps you can guess what's coming next. You do the same thing you did in Section 9.1. If you don't know the value of $\sigma$, you substitute $s$ for $\sigma$ and $t$ for $z$. The test statistic becomes

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

**Example: Mean Weight of Pennies**

The weights of newly minted U.S. pennies are approximately normally distributed and are targeted to have a mean weight of 3.11 g. The random sample of nine pennies shown in Display 9.17 was taken from a production line. Their mean weight is 3.16 g with standard deviation 0.065. What is the value of the test statistic for a test to determine whether the mean has moved away from the target mean?

```
3.06 3.10 3.14 3.18 3.22 3.26
```

Display 9.17 Weights of newly minted pennies.

**Solution**

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{3.16 - 3.11}{0.065/\sqrt{9}} \approx 2.31$$

**Constructing a $t$-Distribution**

Before you can start interpreting the new test statistic $t$, you need to know how it behaves when the null hypothesis is true. If the value of $t$ from your sample fits right into the middle of the distribution of $t$ constructed by assuming the null hypothesis is true, you have no evidence against the null hypothesis. If the value of $t$ from your sample is way out in the tail of the $t$-distribution, you have evidence
against the null hypothesis. In other words, if the value of \( t \) from your sample would be a rare event given the hypothesized mean, be suspicious of that mean. If the value of \( t \) from your sample is a reasonably likely event, the hypothesized mean is plausible.

The following simulation, based on the distribution of IQ scores, will show you how the distribution of \( t \) is different from that of \( z \).

Some IQ tests are constructed so that scores are normally distributed with mean \( \mu = 100 \) and standard deviation \( \sigma = 15 \). Suppose you repeatedly take random samples of size 9 from this distribution and compute the sample mean, \( \bar{x} \). According to the explanation in Section 7.2, the distribution of

\[
    z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{x} - 100}{15/\sqrt{9}}
\]

will be normal even with this small sample size (because the population is normally distributed). The histogram on the left in Display 9.18 shows 1000 values of \( z \), each computed from a random sample of size 9 from the population of IQs.

If you do not know \( \sigma \) (the usual situation), you have to use \( s \) to estimate it. You would then compute

\[
    t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{\bar{x} - 100}{s/\sqrt{9}}
\]

The histogram on the right in Display 9.18 shows 1000 values of \( t \), each computed from a random sample of size 9 from the population of IQs.

Display 9.18  (Left) 1000 values of \( z \) computed from random samples taken from a normally distributed population with mean 100 and standard deviation 15; (right) 1000 values of \( t \) computed from random samples from the same population.

Note that the values of \( t \) are more spread out than are the values of \( z \). That makes sense. Not only does \( \bar{x} \) vary from sample to sample, but \( s \) also varies from sample to sample. And, as you saw in Section 9.1, \( s \) tends to be smaller than \( \sigma \) more often than it tends to be larger than \( \sigma \), making \( t \) larger in absolute value than the corresponding \( z \) more often than it is smaller. Thus, when you compute \( t \), you tend to get more values in the tails of the distribution than when you compute \( z \).
The box summarizes the most important facts about $t$-distributions. The values of $t$ in Table B on page 826 that you used to construct confidence intervals in Section 9.1 come from these distributions.

**The $t$-Distributions**

Suppose you draw random samples of size $n$ from a normally distributed population with mean $\mu$ and unknown standard deviation. The distribution of the values of

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

is called a $t$-distribution. There is a different $t$-distribution for each degree of freedom, $df = n - 1$, where $n$ is the sample size.

A $t$-distribution is mound-shaped, with mean 0 and a spread that depends on the value of $df$. The greater the $df$, the smaller the spread. The spread of any $t$-distribution is greater than that of the standard normal distribution. Display 9.19 shows the $t$-distribution for $df = 4$ plotted on the same graph as the standard normal distribution.

Display 9.19  The $t$-distribution for $df = 4$, and the standard normal distribution.

As the number of degrees of freedom gets larger, the $t$-distribution more closely approximates a normal distribution. In Display 9.20, the $t$-distribution with $df = 9$ is very much like the standard normal distribution.

Display 9.20  The $t$-distribution for $df = 9$, and the standard normal distribution.

You can explore graphs of $t$-distributions using your calculator. [See Calculator Note 9C.]
**P-Values**

Now that you know the distribution of \( t \) when the null hypothesis is true, the next step is to locate where the value of \( t \) computed from your sample lies on this distribution. The standard way to do that is to report a \( P \)-value, sometimes called an observed significance level, just as in Section 8.2. The next example will show you how to find the \( P \)-value and how to interpret it.

**Example: The P-Value for the Mean Weight of Pennies**

As you saw in the example on page 582, the test statistic to see whether the mean weight of newly minted pennies has moved away from the target, 3.11 g, is

\[
 t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{3.16 - 3.11}{0.065/\sqrt{9}} \approx 2.31
\]

Find and interpret the \( P \)-value that corresponds to this value of \( t \).

**Solution**

The \( P \)-value for a two-sided test is the area under the \( t \)-distribution with \( df = 9 - 1 \), or 8, that lies above \( t = 2.31 \) and below \( t = -2.31 \), as shown in Display 9.21.

**Display 9.21**  \( P \)-value for \( t = 2.31 \) on a \( t \)-distribution with \( df = 8 \).

You can get this \( P \)-value from your calculator or software. Display 9.22 shows the \( P \)-value \( 2(0.0248) = 0.0496 \), or approximately 0.05, on various printouts. [See **Calculator Note 9D** to learn how to find a \( P \)-value given \( t \) and \( df \)].

**Display 9.22**  \( P \)-values from Minitab, Fathom, and the TI-84 Plus.
If it is true that the mean has not slipped off the target, 3.11 g, there is only a 0.0496 chance of getting an absolute value of $t$ as large as or even larger than the one from this sample (that is, $t \geq 2.31$ or $t \leq -2.31$). The small $P$-value, 0.0496, indicates that the sample is somewhat inconsistent with the null hypothesis. It looks as if the population mean has moved away from the target.

If you do not have a calculator or software that finds $P$-values, you can get an estimate of the $P$-value from Table B on page 826. The next example shows you how.

**Example: Approximating the $P$-value from Table B**

Suppose you have a two-sided test with $df = 9$ and $t = -3.98$ (see Display 9.23). Approximate the $P$-value for this test using Table B and using a calculator.

![Display 9.23](A test statistic, t, of -3.98.)

**Solution**

Go to Table B on page 826 and find the row with 9 degrees of freedom. Go across the row until you find the absolute value of your value of $t$, which probably will lie between two of the values in the table. The partial $t$-table here shows the two values of $t$ that lie on each side of $|t| = 3.98$.

<table>
<thead>
<tr>
<th>Tail Probability $p$</th>
<th>$df$</th>
<th>.0025</th>
<th>.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>3.690</td>
<td>4.297</td>
<td></td>
</tr>
</tbody>
</table>

The “tail probability” gives the area that lies in each tail. Because you have a two-sided test, you will double the tail probability. If you must use Table B to find a $P$-value, all you can say is that the $P$-value is between $2(0.001)$, or 0.002, and $2(0.0025)$, or 0.005. That is, $0.002 < P$-value $< 0.005$. To find the $P$-value more precisely, use your calculator to get approximately 0.0032.

or

\[ 1 - tcdf(-3.98, 3.98, 9) = 0.0032057987 \]

\[ 2 * tcdf(-1e9, -3.98, 9) = 0.0032057987 \]
As always, the $P$-value is a probability, computed under the assumption that the null hypothesis is true. It tells the chance of seeing a result from a sample that is as extreme as or even more extreme than the one computed from your data when the hypothesized mean is correct.

**Using a $P$-Value in a Test of a Mean**

The $P$-value is the probability of getting a random sample, from a distribution with the mean given in the null hypothesis, that has a value of $t$ that is as extreme as or even more extreme than the observed value of $t$ computed from the sample you observed.

When the $P$-value is close to 0, you have convincing evidence against the null hypothesis. When the $P$-value is large, the result from the sample is consistent with the hypothesized mean and you don’t have evidence against the null hypothesis.

**DISCUSSION**

**$P$-Values**

D9. Study the formula for the test statistic when testing a hypothesis about a population mean, and think about the relationship of the test statistic to the $P$-value.

a. What happens to the $P$-value if the sample standard deviation increases but everything else remains the same?

b. What happens to the $P$-value if the sample size increases but everything else remains the same?

**To Reject or Not to Reject**

$P$-values are very informative about the strength of the evidence against the null hypothesis, but sometimes it is advantageous to know at what level you will reject the null hypothesis even before the data are collected. For what $P$-values will you decide to reset the penny machine? For what $P$-values will you decide to market the new drug? How much evidence will you require before replacing the thermostat? When a yes/no or stop/go decision is needed, a fixed-level test of significance, just as in Section 8.2, has two advantages:

- You decide before collecting data when you will reject the null hypothesis in favor of the alternative hypothesis so that you are not influenced by the data.
- You can control the probability of rejecting a true null hypothesis (Type I error).
When you are asked to do a fixed-level test, you will be given the required **level of significance** (or **significance level** or simply **level**), \( \alpha \), which typically is 5%. To decide whether to reject the null hypothesis in favor of the alternative hypothesis, you compare your \( P \)-value to \( \alpha \).

### Fixed-Level Testing

When you use fixed-level testing, you reject the null hypothesis in favor of the alternative hypothesis if your \( P \)-value is less than the level of significance, \( \alpha \). If your \( P \)-value is greater than or equal to the level of significance, you do not reject the null hypothesis.

The significance level is equal to the probability of rejecting the null hypothesis when it is true (making a Type I error).

The smaller the significance level you choose, the stronger you are requiring the evidence to be in order to reject \( H_0 \). The stronger the evidence you require, the less likely you are to make a Type I error (to reject \( H_0 \) when it is true). However, the stronger the evidence you require, the more likely you are to make a Type II error (to fail to reject \( H_0 \) when it is false).

**Example: Fixed-Level Testing of the Mean Penny Weight**

State the null and alternative hypotheses for a test of significance to determine whether the mean penny weight has moved away from the target, 3.11 g (see the examples on pages 582 and 585). Can you reject the null hypothesis at a level of significance, \( \alpha \), of 0.01? At \( \alpha = 0.05 \)? At \( \alpha = 0.10 \)?

**Solution**

The hypotheses are

\[
H_0: \mu = 3.11, \quad \text{where} \quad \mu \text{ denotes the mean weight of all pennies on this production line}
\]

\[
H_1: \mu \neq 3.11
\]

From the example on page 585, the \( P \)-value for this test is 0.0496.

Because the \( P \)-value is greater than \( \alpha = 0.01 \), you do not reject the null hypothesis at that level of significance. Because the \( P \)-value is less than \( \alpha = 0.05 \), you do reject the null hypothesis at that level of significance. Because the \( P \)-value is less than \( \alpha = 0.10 \), you do reject the null hypothesis at that level of significance.

Note that, in the previous example, if your calculator or software rounded the \( P \)-value to 0.05, you would not have rejected the null hypothesis at a significance level, \( \alpha \), of 0.05. That’s the major drawback of fixed-level testing: It boils down all the data to the two choices “reject” or “don’t reject.” To see what gets lost, think about a court trial where “not guilty” can mean anything from “This guy is so innocent he should never have been brought to trial” to “Everyone on the jury thinks the defendant did it, but the evidence presented isn’t quite strong enough.” In the same way, “do not reject” communicates only a small fraction of the
information in the data. That's why reporting the \( P \)-value has become standard in modern statistical practice.

**The \( t \)-Test**

Tests of significance for means, called **\( t \)-tests**, have the same general structure as significance tests for proportions, although some details are a bit different.

### Components of a Significance Test for a Mean

1. **Name the test and check conditions.** For a test of significance for a mean, the methods of this section require that you check three conditions:
   - **Randomness.** In the case of a survey, the sample must have been randomly selected. In the case of an experiment, the treatments must have been randomly assigned to the experimental units.
   - **Normality.** The sample must look like it's reasonable to assume that it came from a normally distributed population or the sample size must be large enough that the sampling distribution of the sample mean is approximately normal. There is no exact rule for determining whether a normally distributed population is a reasonable assumption or for what constitutes a large enough sample size, but you will learn some guidelines in Section 9.3.
   - **Sample/population size.** In the case of a sample survey, the population size should be at least ten times as large as the sample size.

2. **State your hypotheses.** The null hypothesis is that the population mean, \( \mu \), has a particular value \( \mu_0 \). This is typically abbreviated

   \[
   H_0: \mu = \mu_0
   \]

   The alternative hypothesis can have any one of three forms:

   \[
   H_a: \mu \neq \mu_0 \quad \text{or} \quad H_a: \mu < \mu_0 \quad \text{or} \quad H_a: \mu > \mu_0
   \]

   Be sure to define \( \mu \) in context.

3. **Compute the test statistic, find the \( P \)-value, and draw a sketch.** The test statistic is the distance from the sample mean, \( \bar{x} \), to the hypothesized value, \( \mu_0 \), measured in standard errors:

   \[
   t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}
   \]

   The \( P \)-value given by your calculator or software (or estimated from Table B) is the probability of getting a value of \( t \) that is as extreme as or even more extreme than the one computed from the actual sample *if the null hypothesis is true.*

(continued)
4. Write your conclusion linked to your computations and in the context of the problem. If you are using fixed-level testing, reject the null hypothesis if your $P$-value is less than the level of significance, $\alpha$. If the $P$-value is greater than or equal to $\alpha$, do not reject the null hypothesis. (If you are not given a value of $\alpha$, you can assume that $\alpha$ is 0.05.)

Write a conclusion that relates to the situation and includes an interpretation of your $P$-value.

The next example demonstrates a fixed-level significance test with all four steps included.

**Example: Mass of a Large Bag of Fries**

The statistics class at the University of Wisconsin–Stevens Point (see E5 on page 577) also estimated the average mass of a large bag of french fries at McDonald’s. They bought 30 bags during two different half-hour periods on two consecutive days at the same McDonald’s and weighed the fries. The data are given in Display 9.24. McDonald’s “target value” for the mass of a large order of fries was 171 g. Is there evidence that this McDonald’s wasn’t meeting that target?

![Mass of Large Bag of Fries](chart)

**Display 9.24** Mass of large bags of McDonald’s fries. [Source: Nathan Wetzel, “McDonald’s French Fries. Would You Like Small or Large Fries?” STATS 43 (Spring 2005): 12–14.]
Check conditions and name the test.

State $H_0$ and $H_a$.

Compute the test statistic, find the $P$-value, and draw a sketch.

Write a conclusion linked to your computations and in the context of the situation.

**Solution**

The dot plot in Display 9.24 gives no indication that the large population of masses of large bags of french fries isn’t normally distributed. These certainly aren’t a random sample of the bags of fries, because they were collected on two specific days during two half-hour time periods on each day. The mass of the fries might vary with the person bagging them, and the same person might bag the fries at those times while other people bag the fries on other days or at other times. With that in mind, we will proceed with a one-sample $t$-test for a mean.

The hypotheses are

$H_0: \mu = 171$, where $\mu$ denotes the mean mass, in grams, of all large bags of french fries produced by this McDonald’s

$H_a: \mu \neq 171$

The statistics from the sample are $\bar{x} = 144.07$, $s = 12.28$, and $n = 30$. The test statistic is

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = \frac{144.07 - 171}{12.28 / \sqrt{30}} \approx -12.01$$

The $P$-value from a calculator is $8.85 \times 10^{-13}$.

A $P$-value this close to 0 is very strong evidence against the null hypothesis. A sample that gives a value of $t$ of $-12.01$—or one even more extreme—is almost never going to occur if the mean mass of the bags is 171 g. So you can reject the null hypothesis that the mean mass is 171 g.

Even though this was not a true random sample, the sample of bags collected by these students might be typical of the bags produced during those times or on that day or by that shift of employees. So any generalizations should be made cautiously, with careful attention to describing the population to which the results of the analysis might apply.

You can use your calculator to perform a significance test for a mean. [See Calculator Note 9E.]

In the next example, you will learn one technique for dealing with outliers.

**Example: Distance from the Sun**

“Earth is 93 million miles from the Sun.” You have probably heard that assertion. But is it true? Display 9.25 shows 15 measurements made of the average distance between Earth and the Sun (called the astronomical unit, or A.U.), including Simon Newcomb’s original measurement from 1895. Test the hypothesis that the data are consistent with a true mean A.U. of 93 million miles.

Earth and a total eclipse of the Sun as seen from outer space
Chapter 9 Inference for Means

Astronomical Unit (millions of miles)

| 93.28  |
| 92.83  |
| 92.91  |
| 92.87  |
| 93.00  |
| 92.91  |
| 92.84  |
| 92.98  |
| 92.91  |
| 92.87  |
| 92.88  |
| 92.92  |
| 92.96  |
| 92.96  |
| 92.81  |


<table>
<thead>
<tr>
<th>Astronomical Unit (n = 15), Leaf Unit = 0.010</th>
</tr>
</thead>
<tbody>
<tr>
<td>928</td>
</tr>
<tr>
<td>929</td>
</tr>
<tr>
<td>929</td>
</tr>
<tr>
<td>930</td>
</tr>
<tr>
<td>931</td>
</tr>
<tr>
<td>932</td>
</tr>
</tbody>
</table>

928 | 1 represents 92.81 million miles.

Solution

Check conditions.

This type of problem is commonly referred to as a measurement error problem. There is no population of measurements from which this sample was selected. The measurements were, however, independently determined by different scientists and can be thought of as a “random” sample taken from a conceptual population of all such measurements that could be made.

The stemplot in Display 9.25 shows that one measurement is extremely large compared to the others. This sample does not look as if it could reasonably have been drawn from a normally distributed population. The large measurement, 93.28 million miles, is Newcomb’s original measurement from 1895 and thus is different from the more modern measurements. There might be a good scientific reason to remove it from the data set. So the analysis will be done twice, once with Newcomb’s measurement and once without it.

Letting \( \mu \) denote the true value of the astronomical unit, the hypotheses are

\[ H_0: \mu = 93 \quad \text{and} \quad H_a: \mu \neq 93 \]

From the data, \( \bar{x} = 92.93 \) and \( s = 0.112 \). The test statistic then is given by

\[ t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{92.93 - 93.00}{0.112/\sqrt{15}} \approx -2.42 \]

The P-value is \( \frac{1}{2} P = 0.0148 \) and \( \frac{1}{2} P = 0.0148 \).
From a calculator, the $P$-value is 0.0297.

If it is true that the A.U. is equal to 93 million miles, the chance of getting a mean from a random sample that is as far away as or even farther from 93 million miles as the mean from this sample is only 0.0297. This is convincing evidence against the null hypothesis. If there is no systematic bias in the measurements, you are pretty sure that the astronomical unit differs from 93 million miles.

What about the influence of the outlier? Removing Newcomb’s original 1895 measurement, 93.28 million miles, gives a new observed test statistic of

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{92.90 - 93.00}{0.057/\sqrt{14}} = -6.56$$

The $P$-value is now 0.0000181. This value is even stronger evidence against the hypothesized value, 93 million miles.

Whether the outlier is included or not, you come to the same conclusion: The A.U. is not 93 million miles.

### The t-Test

D10. Compare the lists of components of significance tests for a proportion (page 498) and for a mean (page 589). List all the similarities you find. What are the main differences?

#### The Meaning of “Reject” and “Do Not Reject”

**Large $t$ and Small $P$: Reject the Null Hypothesis**

A test statistic that is large in absolute value tells you that there is a large discrepancy between your sample mean and the hypothesized mean, $\mu_0$. There are three ways you can get a large value of $t$ (and therefore a small $P$-value): The model is wrong, a rare event occurred, or the null hypothesis is false.

- **Wrong model**: The model you used (that you have a random sample from a normally distributed population or a random sample with a large enough sample size) doesn’t fit the data.
- **Rare event**: The model fits and the null hypothesis is true, but just by chance an unlikely event occurred and gave you an unusually large value of the test statistic. In this situation, you will make a *Type I error*, rejecting a true null hypothesis.
- **“Reject” is right**: The model fits, but the true population mean is not $\mu_0$, so you are right to reject $H_0$.

Once you’ve checked the normality of the population as best you can or have a large random sample, you know you have a reasonable model—so either a rare event occurred or the null hypothesis is false. Standard practice is to reject the null hypothesis.
Small $t$ and Large $P$: Do Not Reject the Null Hypothesis

What about a small absolute value of the test statistic? Assuming that your sample fits the conditions for your test, there are two ways you can get a small value of $t$ (and therefore a large $P$-value): The hypothesized mean is right, or the hypothesized mean is wrong but your test doesn’t have enough power to allow you to reject it.

- “Do not reject” is right: The null hypothesis is true, and your sample is consistent with it. You don’t reject the null hypothesis—the correct decision.
- Not enough power: When your sample size is small, there can be a lot of variability in $\bar{x}$. Thus, the sample mean might end up being fairly close to the erroneous mean stated in the null hypothesis. You can’t reject this false null hypothesis, and so you have made a Type II error. This occurs because a significance test’s performance (its power) is measured by its chance of rejecting a hypothesized value. You want the power to be large for null hypothesized values that are not true. As you learned in Section 8.2, increasing your sample size is the best way to get more power in your test.

The Meaning of “Reject” and “Do Not Reject”

D11. Suppose you are doing a fixed-level test at the 1% significance level. Which of these questions can be answered?

A. What is the chance that the test will reject the null hypothesis?
B. If the null hypothesis is true, what is the probability that it will be rejected? That it will not be rejected?
C. If the null hypothesis is false, what is the probability that it will be rejected? That it will not be rejected?

One-Sided Tests

So far in this section, every test has had a two-sided alternative hypothesis: The true mean, $\mu$, is not equal to the standard, $\mu_0$. But there are two other possible alternatives: $\mu$ might be less than the standard, or $\mu$ might be greater than the standard. In real applications, you sometimes can use the context to rule out one of these two possibilities as meaningless, impossible, uninteresting, or irrelevant.

In *Martin v. Westvaco* (Chapter 1), a statistical analysis compared the ages of the workers who were laid off with the ages of workers in the population of employees working for Westvaco at the time of the layoff. An average age for the fired workers that was greater than the population mean would tend to support a claim of age discrimination. On the other hand, the opposite inequality is not relevant: An average age for the fired workers that was less than the overall average would not be evidence of age discrimination, because younger workers aren’t protected under the law.

When the context tells you to use a one-sided alternative hypothesis, your $P$-value is the area on only one side of the $t$-distribution. Such $P$-values are called one-sided or one-tailed $P$-values, and the corresponding test of significance is called a one-sided or one-tailed test, just as in Chapter 8.
Example: One-Sided P-Value

A book about different colleges reports that the mean time students at a particular university study each week is 1015 minutes. A dean says she believes the mean is greater than 1015 minutes. To test her claim, she takes a random sample of 64 students and finds that the sample mean is 1050 minutes, with standard deviation 150 minutes. Is this strong evidence in favor of her claim?

Solution

A random sample of size 64 would produce a nearly normal sampling distribution of the sample mean even if the population were not normally distributed. The university undoubtedly has more than 10(64), or 640, students. This situation requires a one-sided test, because the dean is interested only in the alternative (research) hypothesis that the population mean might be larger than reported.

The hypotheses are

\[ H_0: \mu = 1015, \text{ where } \mu \text{ denotes the mean number of minutes students study each week for all the university’s students} \]

\[ H_a: \mu > 1015 \]

From the data, \( \bar{x} = 1050, s = 150, \) and \( n = 64. \) The test statistic then is given by

\[
t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1050 - 1015}{150/\sqrt{64}} = 1.87
\]

The \( P \)-value is the area to the right of 1.87 under the \( t \)-distribution with \( df = 64 - 1, \) or 63. From a calculator, the \( P \)-value is about 0.0331.

A \( P \)-value of only 0.0331 is fairly strong evidence against the null hypothesis that the mean amount of time spent studying each week is 1015 minutes. Thus, the data support the dean’s claim that the mean is greater than 1015 minutes.

Summary 9.2: A Significance Test for a Mean

A claim is made about the value of a population mean, \( \mu, \) and you want to conduct a test of the significance of this claim by taking a random sample from the population. Then

- the null hypothesis gives the hypothesized, or standard, value of \( \mu \)
- the alternative (research) hypothesis, which can be one- or two-sided, conjectures how the true mean differs from the standard given in the null hypothesis

You must check that you have a random sample that is not more than one-tenth the population size (or a random assignment of treatments to experimental units) and that it is reasonable to assume either that the sample came from a distribution that is approximately normal or that the sample size is large enough (to be explained in Section 9.3).
Just as in constructing a confidence interval for a mean, when you use \( s \) to estimate \( \sigma \), you must use \( t \) rather than \( z \) in doing a significance test for a mean. You can use the test statistic

\[
t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}
\]

Instead of doing a fixed-level test with given significance level \( \alpha \), you can report a \( P \)-value in order to provide more information about the strength of evidence against the null hypothesis than you provide by simply saying “reject” or “don’t reject.” The \( P \)-value gives the probability of seeing a test statistic \( t \) as extreme as or even more extreme than the one computed from the data when the null hypothesis is true. It serves as a measure of the evidence against the null hypothesis. The smaller the \( P \)-value, the greater the evidence in the sample against the null hypothesis. You can get exact \( P \)-values from a calculator or statistical software. You can get approximate \( P \)-values from Table B on page 826.

If you reject a true null hypothesis, you have made a Type I error. If you fail to reject a false null hypothesis, you have made a Type II error. The power of your test depends on its ability to reject the null hypothesis. The larger the sample size and the larger the level of significance, the greater the power of your test to reject a false null hypothesis.

### Practice

**The Test Statistic**

P10. The thermostat in your classroom is set at 72°F, but you think the thermostat isn’t working well. On seven randomly selected days, you measure the temperature at your seat. Your measurements (in degrees Fahrenheit) are 71, 73, 69, 68, 69, 70, and 71. What is the test statistic for a significance test of whether the mean temperature at your seat is different from 72°F?

P11. In 2003, Pell Grants totaling $166,733,957 were awarded to 76,525 Minnesota college students. You believe that the mean amount has changed since 2003. You take a random sample of 35 Minnesota college students who currently receive Pell Grants and get a mean amount of $2317.14, with a standard deviation of $754.00. What is the test statistic for a test to determine whether the mean Pell Grant amount for Minnesota college students has changed since 2003? [Source: www.ohe.state.mn.us.]

**Constructing a \( t \)-Distribution**

P12. One of the distributions in Display 9.26 is a \( t \)-distribution, and the other is the standard normal distribution. Which is which? Explain how you know.

Display 9.26  A \( t \)-distribution and a standard normal distribution.

P13. You plan to use simulation to construct an approximate \( t \)-distribution with 6 degrees of freedom by taking random samples from a population of IQ scores that are normally distributed with mean, \( \mu \), 100 and standard deviation, \( \sigma \), 15.
a. Describe how you will do one run of this simulation.

b. Produce three values of $t$ using your simulation.

**P-Values**

P14. In P10, you computed the test statistic for testing that the mean temperature at your seat isn’t 72°F. Find and interpret the $P$-value.

P15. In P11, you computed the test statistic for testing whether the mean Pell Grant amount for Minnesota college students has changed since 2003. Find and interpret the $P$-value.

**To Reject or Not to Reject**

P16. In P14, you found the $P$-value for testing that the mean temperature at your seat isn’t 72°F.

a. State the hypotheses.

b. Would you reject the null hypothesis at the 10% level? The 5% level? The 1% level?

P17. In P15, you found the $P$-value for testing whether the mean Pell Grant in Minnesota has changed since 2003.

a. State the hypotheses.

b. Would you reject the null hypothesis at the 10% level? The 5% level? The 1% level?

P18. You get the SAT scores of a random sample of nine students from State University and calculate a sample mean of 1535 with standard deviation 250. A university official states that the mean SAT score for students at the school is 1700. You are asked to validate his claim.

a. State the hypotheses.

b. Would you reject the hypothesis that State University has a mean SAT score of 1700 at the 10% level? The 5% level? The 1% level?

**The $t$-Test**

P19. For the aldrin data of P1 on page 573, carry out a test of whether the true mean differs from 4 nanograms per liter. Use $\alpha = 0.05$.

P20. The question about mean body temperature addressed with the data in Display 9.3 on page 567 could have been phrased differently: “Does the mean body temperature differ from 98.6°F?” Do these data provide convincing evidence that the population mean for males differs from 98.6°F? Answer the same question for the female population. State your conclusions in terms of $P$-values.

**The Meaning of “Reject” and “Do Not Reject”**

P21. Return to the significance tests indicated and tell whether a Type I error, a Type II error, both errors, or neither error might have been made.

a. french fries example on pages 590–591

b. astronomical unit example on pages 591–593

c. aldrin, P19

P22. In each situation below, assume the null hypothesis is actually false. Tell which test will have more power, all else being equal.

a. a test with $n = 12$; a test with $n = 16$

b. a test with $\alpha = 0.05$; a test with $\alpha = 0.01$

**One-Sided Tests**

P23. The board of the Gainesville Home Builders’ Association knows the selling price of all new houses sold in the city last month and believes that the mean selling price has gone up this month. To check this claim, the board obtains a random sample of the selling prices of houses sold this month.

a. Should the alternative hypothesis be one-sided or two-sided?

b. Tell what $\bar{x}$ and $\mu$ are for the problem.

c. State the null and alternative hypotheses.

P24. The situation is the same as in P10 on page 596. The thermostat in your classroom is set at 72°F, but this time, before you collect your data, you are convinced that the room tends to be colder. On seven randomly selected days, you measure the temperature at your seat. Your measurements are 71, 73, 69, 68, 69, 70, and 71. Perform all four steps of a test of whether the mean temperature at your seat is lower than 72°F.
Exercises

E17. In E3 on page 576, Jack and Jill opened a water-bottling factory. The distribution of the number of ounces of water in the bottles is approximately normal. The mean, μ, is supposed to be 16 oz, but the water-filling machine slips away from that amount occasionally and has to be readjusted. Jack and Jill take a random sample of ten bottles from today’s production and weigh the water in each. The weights (in ounces) are

15.91 16.08 16.08 15.94 16.02
15.94 15.96 16.03 15.82 15.96

Should Jack and Jill readjust the machine? Do all four steps in a test of significance, writing the conclusion in terms of the P-value.

E18. In E5 on page 577, a statistics class decided to check the weights of bags of small fries at a local McDonald’s to see if, on average, they met the “target value” of 74 g. They bought 32 bags during two different time periods on two consecutive days and weighed the fries. The data are given in Display 9.12 on page 577. Is there evidence that this McDonald’s wasn’t meeting its target? Do all four steps in a test of significance, writing the conclusion in terms of the P-value.

E19. Fifteen students were given pieces of paper with five vertical line segments and asked to mark the midpoint of each segment. Then, using a ruler, each student measured how far each mark was from the real midpoint of each segment. The students then each averaged their five errors. If the average is positive, the student tended to place the midpoint too high. If the average is negative, the student tended to place the midpoint too low. The results are given in Display 9.27. Is there evidence of statistically significant measurement bias? That is, on average, do students tend to place the midpoints either too high or too low? Do all four steps in a test of significance, writing the conclusion in terms of the P-value.

E20. Refer to E19. Fourteen different students were given the same instructions, but with pieces of paper with five horizontal line segments. The results are given in Display 9.28. Is there evidence of statistically significant measurement bias? That is, on average, do students tend to place the midpoints either too far to the right or too far to the left? Do all four steps in a test of significance, writing the conclusion in terms of the P-value.
E22. At sea level, water boils at 212°F. Theory says that, at high altitude, water boils at a lower temperature than it does at sea level. You plan to test this theory by flying to the mile-high Denver airport and observing the boiling point of water there.

a. Should the alternative hypothesis be one-sided or two-sided?

b. Tell what \( \bar{x} \) and \( \mu \) are for the problem.

c. State the null and alternative hypotheses.

E23. Munchie's Potato Chip Company claims that the weight of the contents of a 10-oz bag of chips is normally distributed, with mean 10 oz. A consumer group, Snack Munchers for Truth (SMFT), says that the average weight is less than this. SMFT weighs the contents of 15 randomly selected bags of potato chips and gets the weights in Display 9.29. Test SMFT’s claim, including all four steps of a significance test and writing the conclusion in terms of the P-value.

E24. A sample of 15 monthly rents for two-bedroom apartments was selected from a recent edition of the Gainesville Sun. The 15 values are symmetric, with no outliers.

Display 9.28  Average error in estimating the location of the midpoints of horizontal line segments. (A positive value indicates that the average estimate was too far to the right.)

Display 9.29  Weights of bags of chips.
The mean of this sample is $636 with standard deviation $121.

a. A group of students thinks the mean rent for two-bedroom apartments is at least $650. Test their claim by doing a test of significance, writing the conclusion in terms of the $P$-value.

b. A newspaper reports that the average monthly rent for two-bedroom apartments last year was $500. Is there statistical evidence of a change in mean rent from last year to this year? Test this claim by doing a test of significance, writing the conclusion in terms of the $P$-value.

E25. The heights of young women in the United States are approximately normally distributed, with mean 64.8 in. The heights of the 11 players on a recent roster of the WNBA Chicago Sky basketball team are (in inches) 72, 74, 76, 71, 76, 68, 69, 76, 71, 73, and 70. [Source: www.wnba.com.] Is there sufficient evidence to say that the mean height of these 11 players is so much larger than the population mean that the difference cannot reasonably be attributed to chance alone? (You might want to refer to “Inference for an Observational Study” in Section 8.5.)

E26. The heights of young men in the United States are approximately normally distributed, with mean 70.1 in. The heights of the young men in the Hamilton family are (in inches) 71, 72, 71, 72, and 70.

a. Is there sufficient evidence to say that the difference between the mean height of these family members and the population mean cannot reasonably be attributed to chance alone? (You might want to refer to “Inference for an Observational Study” in Section 8.5.)

b. Has the null hypothesis been rejected primarily because the sample size is so small, because the mean is so close to that of all young men, or because the variability is so small in the Hamilton family?

E27. Return to the significance test in E17. Rewrite the conclusion for a fixed-level test with $\alpha$ equal to 0.10.

E28. Return to the significance test in E18. Rewrite the conclusion for a fixed-level test with $\alpha$ equal to 0.04.

E29. Return to the significance test in E23.

a. Rewrite the conclusion for a fixed-level test at the 0.05 level.

b. In fact, these values were selected at random from a normal distribution with mean 9.9. Has an error been made? If so, which type?

c. Could you have reached the conclusion in part a based only on a confidence interval? Explain.

E30. Return to the significance test in E24, part b.

a. Rewrite the conclusion for a fixed-level test with $\alpha$ equal to 0.02.

b. Could you have reached the conclusion in part a based only on a confidence interval? Explain.

E31. Select the best answer. A $P$-value measures

A. the probability that the null hypothesis is true
B. the probability that the null hypothesis is false
C. the probability that an alternative hypothesis is true
D. the probability of seeing a value of $t$ at least as extreme as the one observed, given that the null hypothesis is true

E32. Select each correct description of the power of a test.

A. gives the probability of rejecting the null hypothesis
B. gives the probability of failing to reject the null hypothesis
C. increases by increasing the sample size
D. increases by increasing the level of significance
E. isn't an issue if the null hypothesis is true
E33. Suppose you have a sample of size 23 and the value of your test statistic is 1.645.

a. Draw a sketch of a *t*-distribution and shade the area that corresponds to the *P*-value for a two-sided test.

b. Identify the part of the shaded area that corresponds to evidence that \( \mu > \mu_0 \).

What is the relationship between the two-sided and one-sided *P*-values for a given value of the test statistic?

c. If all else is equal and the alternative hypothesis is in the right direction, will the *P*-value be larger for a one-sided test or a two-sided test?

E34. For the aldrin data of P1 on page 573, find the *P*-value for each alternative hypothesis.

a. \( H_a: \mu \neq 4 \)

b. \( H_a: \mu > 4 \)

c. \( H_a: \mu < 4 \)

E35. Suppose \( H_0 \) is true and you are using a significance level of 5%. Which gives a larger chance of a Type I error, a one-sided test or a two-sided test? Explain.

E36. If the null hypothesis is true, does using \( \alpha = 0.05 \) or \( \alpha = 0.01 \) give a larger chance of a Type I error?

E37. In each part, the null hypothesis is actually false. Which test will have more power, all other things being equal?

a. a test with \( \alpha = 0.01 \); a test with \( \alpha = 0.10 \)

b. a test with \( n = 45 \); a test with \( n = 29 \)

c. a one-sided test; a two-sided test

E38. Suppose \( \mu > 0 \) and you are using a significance level of 5%. Which of these three tests has the greatest power (probability of rejecting the false null hypothesis that \( \mu = 0 \))?

A. a one-sided test of \( H_0 \) versus \( H_a: \mu > 0 \)

B. a two-sided test

C. a one-sided test of \( H_0 \) versus \( H_a: \mu < 0 \)

E39. Histograms A and B in Display 9.30 were generated from 200 random samples of size 4, each selected from a normal distribution with \( \mu = 100 \) and \( \sigma = 20 \). One histogram shows the 200 *z*-values (one for each sample, using \( \sigma \)) for testing the hypothesis that the population mean is, in fact, 100, and the other shows the 200 *t*-values (using \( s \)).

Display 9.30 Histograms of the *t*-values and *z*-values computed for 200 random samples of size 4 taken from a normal distribution with mean 100 and standard deviation 20.

a. Compare the shapes, centers, and spreads of these two distributions.

b. Choose which is the distribution of *t*-values, and give the reason for your choice.

E40. Degrees of freedom tell how much information in your sample is available for estimating the standard deviation of the population. In a *t*-test, you use \( s \) to estimate \( \sigma \). The more deviations from the mean, \( (x - \overline{x}) \), you have available to use in the formula for \( s \), the closer \( s \) should be to \( \sigma \).

But, as you will see in this exercise, not all of these deviations give you independent information.

a. For a sample of size \( n \), how many deviations from the mean are there? What is their sum? (See Section 2.3.)

b. Suppose you have a random sample of size \( n = 1 \) from a completely unknown population. Call the sample value \( x_1 \). Does the value of \( x_1 \), by itself, give you information about the spread of the population?

c. Next suppose you have a sample of size \( n = 2 \), with values \( x_1 \) and \( x_2 \). If one deviation, \( x_1 - \overline{x} \), is 3, what is the other deviation?

d. Now suppose you have a sample of size \( n = 3 \). Suppose you know two of the
deviations, \( x_1 - \bar{x} = 3 \) and \( x_2 - \bar{x} = -1 \). What is the third deviation?

e. Finally, suppose you have a sample of size \( n \). Show how to find the final deviation once you know all the others. With a sample of size \( n \), how many deviations give you independent information about the size of \( \sigma \) and how many are redundant?

f. Explain what your answers to parts a–e have to do with degrees of freedom.

## 9.3 When Things Aren’t Normal

If you think about what you’ve done so far in terms of “the big three” for distributions—center, spread, and shape—you’ll see that there’s more work to do.

1. **Center:** The population mean is the goal of our inference. You don’t know the mean, but that’s the whole point: If you knew the mean, you wouldn’t be trying to estimate it.

2. **Spread:** In real applications, you almost never know the population standard deviation in confidence intervals and significance tests, so you have to use \( s \) as an estimate of \( \sigma \). In Section 9.1, you saw that, for small samples, using \( s \) in place of \( \sigma \) has a pretty big effect on confidence intervals. Unless you make a suitable adjustment to the width of the interval by using \( t^* \) instead of \( z^* \), its capture rate will be way off.

3. **Shape:** “Officially,” inference using the \( t \)-distribution requires a normally distributed population. But official rules don’t always have to be followed exactly. You can safely get away with less than a normally distributed population, because the \( t \)-procedures are robust (not very sensitive) to departures from normality. You’ll still have to check a condition about shape, which will be described on pages 607–608.

First, you will see the reason for the condition requiring a normal distribution by investigating what happens to the capture rate when small samples are taken from a skewed population.

### The Effect of Skewness and Outliers

In the examples so far, the shape of the population has been approximately normally distributed or the sample size has been large, but often that’s not the case. Display 9.31 shows the brain weights of 68 species of animals used in a study of sleep. These 68 species will serve as the population. This distribution is highly skewed toward the larger values. In Activity 9.3a, you will observe the effect the skewness has on the capture rate of confidence intervals.
### Display 9.31 Brain weights for a selection of species.

ACTIVITY 9.3a

The Effect of Skewness on Confidence Intervals

**What you’ll need:** the data set of Display 9.31

Your goal is to see what happens to the capture rate when you construct many confidence intervals based on random samples taken from the population of brain weights in Display 9.31.

1. Decide on a way to get a random sample of five of the species. Select the sample and record the brain weights of the five species in your sample.
2. Construct a 95% confidence interval estimate of the mean brain weight of all the species using the sample mean, the sample standard deviation, and the appropriate value of $t^*$.
3. Does your confidence interval include the true mean weight, 394.49 g? If not, does the interval lie below the true mean or above it?
4. Repeat steps 1 through 3 with a second sample of size 5.
5. Combine your answers with those of other groups in the class to compute an estimate of the overall capture rate. Is it close to 95%?
6. Looking at the capture rate from your simulation and at the locations of the intervals that do not capture the population mean, write a brief statement about what happens to confidence intervals when techniques based on the normal distribution are used with distributions that are highly skewed toward large values.

The lesson to be learned from Activity 9.3a is that if you have data from a distribution that is far from normal, confidence intervals and hypothesis tests might not behave the way they are supposed to.

**The Effect of Skewness and Outliers**

If the sample size is small and if the underlying distribution is highly skewed rather than normal or has extreme outliers, then the capture rate for an interval of the form $\bar{x} \pm t^* \cdot s/\sqrt{n}$ might be substantially lower than the advertised capture rate, for example, 95%, and a test of significance based on the normal distribution will falsely reject $H_0$ substantially more often than the advertised rate, for example, 5%.

**What If My Population Is Not Normal?**

What can you do about skewness and outliers? Most important, before you do any tests or construct any intervals, always plot your data to see their shape. If the plot looks as if the data came from a normal distribution, you don’t need to worry about shape. On the other hand, if your plot shows any major deviations from a normal shape, you will have to decide whether to try one of the three approaches discussed next.
I. Try a Transformation.

Fortunately, skewed distributions almost always can be made much more nearly symmetric by transforming them to a new scale. If you can find a change of scale that makes your data look roughly normal, once again you don't have to worry about shape. Sometimes a change of scale will also take care of what at first looked liked outliers, making them more like “just part of the herd.” [See Calculator Note 3J to review how to perform shape-changing transformations.]

### Example: The Log Transformation

Take the logarithm of each brain weight in Display 9.31. Is it now safe to proceed with a confidence interval based on $t$?

**Solution**

Display 9.32 shows a histogram of the brain weights after taking the natural logarithm of each value. Notice how the outliers have been drawn in toward the center of the data in a way that makes the distribution much more nearly symmetric. This distribution still is not normal, but it is symmetric enough to make the distribution of the sample means for 100 samples of size 5 in Display 9.33 look nearly symmetric. Thus, the confidence interval using $t$ will work well on the transformed data.

![Display 9.32: Logarithms of brain weights. Display 9.33: Sample means for 100 samples of size 5 from the logarithms of brain weights.]

### Example: The Reciprocal Transformation

Gasoline mileage for cars typically is computed in miles per gallon. But why not use gallons per mile? Display 9.34 shows estimated city miles per gallon (mi/gal) for a random sample of compact car models, along with the gallons per mile (gal/mi). Which distribution is better suited for inference? Construct and interpret a 95% confidence interval for the population mean of that distribution.
### Table

<table>
<thead>
<tr>
<th>Manufacturer</th>
<th>Model</th>
<th>mi/gal</th>
<th>gal/mi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volkswagen</td>
<td>Golf M-5</td>
<td>37</td>
<td>0.027</td>
</tr>
<tr>
<td>Kia</td>
<td>Rio M-5</td>
<td>32</td>
<td>0.031</td>
</tr>
<tr>
<td>Toyota</td>
<td>Corolla M-5</td>
<td>32</td>
<td>0.031</td>
</tr>
<tr>
<td>Mitsubishi</td>
<td>Lancer M-5</td>
<td>27</td>
<td>0.037</td>
</tr>
<tr>
<td>Ford</td>
<td>Focus FWD M-5</td>
<td>26</td>
<td>0.038</td>
</tr>
<tr>
<td>Mazda</td>
<td>3 A-S5</td>
<td>25</td>
<td>0.040</td>
</tr>
<tr>
<td>Suzuki</td>
<td>Aerio M-5</td>
<td>25</td>
<td>0.040</td>
</tr>
<tr>
<td>Saturn</td>
<td>Ion A-4</td>
<td>24</td>
<td>0.042</td>
</tr>
<tr>
<td>Mitsubishi</td>
<td>Lancer M-5</td>
<td>23</td>
<td>0.043</td>
</tr>
<tr>
<td>Volkswagen</td>
<td>Jetta M-6</td>
<td>23</td>
<td>0.043</td>
</tr>
<tr>
<td>Nissan</td>
<td>Sentra A-4</td>
<td>23</td>
<td>0.043</td>
</tr>
<tr>
<td>Chevrolet</td>
<td>Optra 5 A-4</td>
<td>22</td>
<td>0.045</td>
</tr>
<tr>
<td>Subaru</td>
<td>Legacy AWD M-5</td>
<td>22</td>
<td>0.045</td>
</tr>
<tr>
<td>Volvo</td>
<td>S40AWD M-6</td>
<td>20</td>
<td>0.050</td>
</tr>
<tr>
<td>BMW</td>
<td>325I M-6</td>
<td>20</td>
<td>0.050</td>
</tr>
</tbody>
</table>

### Display 9.34


### Solution

The boxplots show that the gallons per mile distribution doesn’t have an outlier and is less skewed than the miles per gallon distribution and thus should be better suited to inference.

The second condition for constructing a confidence interval for a mean also is met—you were told this is a random sample taken from the population of compact car models. The third condition is met if there are at least 10(15), or 150, models of compact cars. Because this isn’t the case, the confidence interval will be wider than necessary for 95% confidence.

The statistics from the sample are \( \bar{x} = 0.0405 \), \( s = 0.00669 \), and \( n = 15 \). For \( df = 15 - 1 = 14 \), use \( t^* = 2.145 \). A 95% confidence interval for the sample mean is

\[
\bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}} = 0.0405 \pm 2.145 \cdot \frac{0.00669}{\sqrt{15}} = 0.0405 \pm 0.0037
\]

You are 95% confident that the mean gallons per mile of all models of compact cars lies in the interval from 0.0368 to 0.0442. So, any mean gal/mi in that interval could have produced the result from the sample as a reasonably likely outcome.

Applying the methodology of this chapter to gal/mi measurements will, over many random samples, produce intervals with a capture rate closer to the nominal 95% than would using miles per gallon measurements.
One lesson of the previous example is that whenever data come in the form of a ratio, think about what might happen if you inverted the ratio. Similarly, a rate like “30 customers are served per hour” might just as naturally be expressed as “2 minutes per customer.”

II. Do the Analysis With and Without the Outliers.

If changing the scale doesn’t take care of the outliers, do two parallel analyses, one with all the data and the other with the outliers removed. That’s what was done in the example on pages 591–593. If both analyses lead to the same conclusion, you’re all set. But if the two conclusions differ, you need more data: You don’t want your conclusion to depend on what you assume about one or two observations!

III. Get a Large Sample Size.

The worst cases are small samples with extreme skewness or with extreme outliers. For moderate sample sizes, you can rely on the robustness of the $t$-procedures: The $t$-procedures are comparatively insensitive to departures from normality, especially for larger samples. What this means in practice is that the true capture rates and significance levels will be close to the advertised values except in extreme situations. For large samples, the value of $s$ won’t vary much from sample to sample and the Central Limit Theorem says that the sampling distribution of $\bar{x}$ will be approximately normal, so it is safe to use the $t$-procedure.

As a rough guide, you can rely on the 15/40 rule, described in the box.

### 15/40 Guideline for Inference Using $t$-Procedures

First, plot your data. A boxplot is especially helpful in identifying shape and outliers. If your random sample looks as if it reasonably could have come from a normal distribution, you can proceed. If you suspect that the data didn’t come from a normal distribution, follow the guideline for your sample size.

- **If your sample size is less than 15:** Be very careful. Your data or transformed data must look as if they came from a normal distribution—little skewness, no outliers.

- **If your sample size is between 15 and 40:** Proceed with caution. Strongly skewed distributions should be transformed to a scale that makes them more nearly symmetric before you use a $t$-procedure. If you have extreme outliers, a transformation might be in order. If you don’t transform the data or if the outliers remain even after a change of scale, do two versions of your significance test or confidence interval, one with and one without the outliers. Don’t rely on any conclusions that depend on whether you include the outliers.

- **If your sample size is over 40:** You’re in good shape. Your sample size is large enough that skewness will not reduce capture rates or alter significance levels enough to matter. Still, if your sample shows strong skewness, it is worth asking whether a change of scale would make the usual summary

(continued)
statistics (especially the standard deviation) more meaningful. Even though outliers might not have much effect on capture rates or significance levels, you still should check by doing two versions of your $t$-procedure.

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Small samples: Inferences are quite sensitive to shape.</th>
<th>Moderate-sized samples: Transform if data are strongly skewed.</th>
<th>Large samples: Transforming for symmetry is less important.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Display 9.35  The 15/40 guideline for inference based on the $t$-distribution.

What If My Population Is Not Normal?

D12. Discuss whether a transformation might be necessary in order to estimate the population mean in each scenario.
   a. You are taking a sample of size 10 from the population of prices of single-family houses in your city.
   b. You are taking a sample of size 100 from the population of prices of single-family houses in your city.
   c. You are taking a sample of size 5 from the SAT scores of students entering a college as freshmen.
   d. You are taking a sample of size 20 from the waiting times of customers at a bank drive-up window.

Summary 9.3: When Things Aren’t Normal

Small samples from a skewed population will tend to produce confidence intervals that are too narrow and off-center. The capture rate of such intervals will be less than the confidence level. If the sample size is small, be careful! If the sample size is large, however, the skewness of the population is of little consequence. If the sample size is moderate, then the sampling distribution might be slightly skewed, but $t$-procedures are robust to slight departures from normality; that is, they still will work well.

Always plot your data first and then proceed with a confidence interval or test of significance using the 15/40 guideline:
   • If the sample size is less than 15, make sure your data or transformed data have little skewness and no outliers.
   • If the sample size is between 15 and 40, your data or transformed data should be nearly symmetric, with no extreme outliers. (If you don’t transform the data or if the outliers remain even after a change of scale, do two versions
of your significance test or confidence interval, one with and one without the outliers. Don’t rely on any conclusions that depend on whether you include the outliers.)

- If the sample size is over 40, it’s safe to proceed. If your sample shows strong skewness, it still is worth asking whether a change of scale would make the usual summary statistics (especially the standard deviation) more meaningful.

With small samples, in addition to constructing plots, think carefully about the nature of the population before concluding that approximate normality is a reasonable assumption.

### Practice

**What If My Population Is Not Normal?**

P25. Pretend that each data set (A–D) is a random sample and that you want to do a significance test or construct a confidence interval for the unknown mean. Use the sample size and the shape of the distribution to decide which description (I–IV) best fits each data set.

A. weights of bears (Display 2.6 on page 33)
B. number of passengers at airports (Display 7.53 on page 462)
C. speed of mammals (Display 2.25 on page 44)
D. female life expectancy in Africa and Europe (Display 2.53 on page 69)

I. There are no outliers, and there is no evidence of skewness. Methods based on the normal distribution are suitable.

II. The distribution is not symmetric, but the sample is large enough that it is reasonable to rely on the robustness of the \( t \)-procedure and construct a confidence interval, without transforming the data to a new scale.

III. The shape suggests transforming. With a larger sample, this might not be necessary, but for a skewed sample of this size transforming is worth trying.

IV. It would be a good idea to analyze this data set twice, once with the outliers and once without.

P26. Display 9.36 shows the number of fries in 30 large bags, similar to the example on page 590. Suppose you want to test that the mean number of fries in a large bag is 90.

<table>
<thead>
<tr>
<th>Number of Fries in Large Bag</th>
<th>Number of Fries in Large Bag</th>
</tr>
</thead>
<tbody>
<tr>
<td>63</td>
<td>108</td>
</tr>
<tr>
<td>122</td>
<td>92</td>
</tr>
<tr>
<td>92</td>
<td>67</td>
</tr>
<tr>
<td>96</td>
<td>90</td>
</tr>
<tr>
<td>72</td>
<td>103</td>
</tr>
<tr>
<td>68</td>
<td>72</td>
</tr>
<tr>
<td>88</td>
<td>76</td>
</tr>
<tr>
<td>64</td>
<td>93</td>
</tr>
<tr>
<td>96</td>
<td>93</td>
</tr>
<tr>
<td>76</td>
<td>99</td>
</tr>
<tr>
<td>111</td>
<td>67</td>
</tr>
<tr>
<td>144</td>
<td>101</td>
</tr>
<tr>
<td>93</td>
<td>92</td>
</tr>
<tr>
<td>118</td>
<td>110</td>
</tr>
<tr>
<td>79</td>
<td>94</td>
</tr>
</tbody>
</table>

Display 9.36  Number of fries in a large bag. [Source: Nathan Wetzel, “McDonald’s French Fries. Would You Like Small or Large Fries?” STATS 43 (Spring 2005): 12–14.]

a. How would you suggest handling the analysis in this case? Give your reasons.
b. Do your analysis and report your results.
Exercises

For E41 and E42: Pretend that each data set described is a random sample and that you want to do a significance test or construct a confidence interval for the unknown mean. Use the sample size and the shape of the distribution to decide which of these descriptions (I–IV) best fits each data set.

I. There are no outliers, and there is no evidence of skewness. Methods based on the normal distribution are suitable.

II. The distribution is not symmetric, but the sample is large enough that it is reasonable to rely on the robustness of the $t$-procedure and construct a confidence interval, without transforming the data to a new scale.

III. The shape suggests transforming. With a larger sample, this might not be necessary, but for a skewed sample of this size transforming is worth trying.

IV. It would be a good idea to analyze this data set twice, once with the outliers and once without.

E41. See the instructions above.

A. weights, in ounces, of bags of potato chips

B. per capita gross domestic product (GNP) for various countries

C. batting averages of American League players

D. self-reported grade-point averages of 67 students

E42. See the instructions above.

A. mean number of people per room for various countries

B. record low temperatures of national capitals

C. student errors in estimating the midpoint of a segment
E43. Health insurance companies look for ways to lower costs, and one way is to shorten the length of hospital stays. A study to compare two large insurance companies on length of stay (LOS) for pediatric asthma patients randomly sampled 393 cases from Insurer A. Summary statistics and a histogram of the data are shown in Display 9.37.


- Is it appropriate to construct a confidence interval without transforming these data? Explain.
- Regardless of your answer to part a, estimate the mean LOS for Insurer A in a 90% confidence interval.
- Would you be more concerned about constructing a confidence interval without a transformation if the sample size was 40 instead of nearly 400? What about a sample size of 4 instead of nearly 400?
- Compare your confidence intervals in part b of E43 and E44. Does it appear that the two insurers differ in mean LOS?

E44. An independent random sample of 396 cases from Insurer B gave the results for length of stay summarized in Display 9.38.

**Display 9.38** Summary and data plot for lengths of hospital stays for Insurer B.

- Is it appropriate to construct a confidence interval without transforming these data? Explain.
- Regardless of your answer to part a, estimate the mean LOS for Insurer B in a 90% confidence interval.
- Would you be more concerned about constructing a confidence interval without a transformation if the sample size was 40 instead of nearly 400? What about a sample size of 4 instead of nearly 400?
- Compare your confidence intervals in part b of E43 and E44. Does it appear that the two insurers differ in mean LOS?

E45. Display 9.39 (on the next page) gives the location and mass of 31 black holes. Assume that these can be considered a random sample of all black holes in the universe.

- Have the conditions been met for constructing a confidence interval for a mean? If not, how would you recommend that the analysis proceed?
- Construct and interpret a 90% confidence interval for a mean, proceeding in the way you recommended in part a.
Galaxy | Mass (in 100,000 solar units)
--- | ---
Milky Way | 18
N221 = M32 | 25
N224 = M31 | 450
N821 | 370
N1023 | 440
N1068 | 150
N2778 | 140
N2787 | 410
N3115 | 10,000
N3245 | 2,100
N3377 | 1,000
N3379 | 1,000
N3384 | 160
N3608 | 1,900
N4258 | 390
N4261 | 5,200
N4291 | 3,100
N4342 | 3,000
N4459 | 700
N4473 | 1,100
N4486 = M87 | 30,000
N4564 | 560
N4596 | 780
N4649 | 20,000
N4697 | 1,700
N4742 | 140
N5845 | 2,400
N6251 | 5,300
N7052 | 3,300
N7457 | 35
IC 1459 | 25,000

E46. The ages, in years, of a random sample of 16 pennies are 1, 1, 1, 1, 1, 1, 2, 2, 3, 4, 4, 11, 11, 12, 26, and 30. You want to test that the mean penny age is 7 years.

a. The boxplot in Display 9.40 shows that 30 is an outlier. Should it be removed from the data set? Explain why or why not.

Display 9.40 Boxplot of the ages of 16 pennies.

b. Are the conditions met for a significance test of the mean age of all pennies? If not, what transformation would you suggest? Give your reason.

c. Try your transformation. Are the conditions better for inference?

E47. Look again at the brain weight data in Display 9.31. This time, consider it a random sample taken from all animal species.

a. If you omit the three obvious outliers, you get the boxplot in Display 9.41. Has removing the three outliers improved the situation?

Display 9.41 Boxplot of brain weights with outliers removed.

b. Construct the 95% confidence interval for the mean using the set of all 68 species. The summary statistics are $\bar{x} = 394.5$ and $s = 1207.0$. Then construct the 95% confidence interval for the mean with the three original outliers removed. The summary statistics are $\bar{x} \approx 149.4$ and $s \approx 323.2$. 

612 Chapter 9 Inference for Means
c. How do the centers of the two confidence intervals in part b compare? The widths? What does this tell you about the sensitivity of confidence intervals to outliers?

d. If you take the natural log of each of the original 68 brain weights, you get the boxplot in Display 9.42 and these summary statistics: $\bar{x} \approx 2.977$ and $s \approx 2.686$. Construct and interpret a 95% confidence interval for the population mean on this transformed scale.

Display 9.42  Boxplot of $\ln(\text{brain weights})$.

E48. Fish absorb mercury as water passes through their gills, and too much mercury makes the fish unfit for human consumption. In 1994, the state of Maine issued a health advisory warning that people should be careful about eating fish from Maine lakes because of the high levels of mercury. Before the warning, data on the status of Maine lakes were collected by the U.S. Environmental Protection Agency (EPA) working with the state. Fish were taken from a random sample of lakes, and their mercury content was measured in parts per million (ppm). Display 9.43 shows a subset of those data from a random sample of 35 lakes.

a. Are there any problems with computing a confidence interval estimate of the mean mercury levels in fish for the lakes of Maine using the values in the table? If so, how do you recommend handling the analysis?

b. One newspaper headline proclaimed, “Mercury: Maine Fish Are Contaminated by This Deadly Poison.” Most states consider mercury levels of 0.5 ppm as the borderline for issuing a health advisory. Does it appear that, on average, Maine lakes deserved the headline? Justify your answer statistically.

<table>
<thead>
<tr>
<th>Mercury (ppm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.05 1.22 0.77 0.25 0.4 0.47 0.94</td>
</tr>
<tr>
<td>0.23 0.24 0.67 0.13 0.45 0.37 0.36</td>
</tr>
<tr>
<td>0.1 0.9 0.6 0.29 1.12 0.29 0.54</td>
</tr>
<tr>
<td>0.77 2.5 0.68 0.41 0.32 0.43 0.86</td>
</tr>
<tr>
<td>0.91 0.34 0.22 0.21 0.37 0.16 0.49</td>
</tr>
</tbody>
</table>

Display 9.43  Mercury content of fish in Maine lakes.  

E49. Almost all of the rain in Los Angeles falls in the winter, so a season of rainfall is measured from July 1 through the end of June. Display 9.44 gives a summary of the rainfall in Los Angeles for the last 128 seasons.

a. You want to estimate mean seasonal rainfall in a confidence interval and can afford to check a random sample of only seven seasons. If you suspect that the population has a shape like that in the histogram in Display 9.44, can you proceed, or should you consider a transformation?

b. Display 9.45 shows a reciprocal transformation of the rainfall data. The
mean of the distribution is about 0.081. Does this mean have a meaningful interpretation? Are you satisfied with the conditions for inference now?

Display 9.45  Los Angeles rainfall, in years per inch, for 128 seasons.

c. Display 9.46 shows a log transformation of the rainfall data. Does this transformation do a better or worse job than the reciprocal transformation in satisfying the conditions for inference?

Display 9.46  Log transformation of inches per season of Los Angeles rainfall.

d. If, instead of taking a random sample of size 7, you were able to consider these 128 seasons to be a random sample of Los Angeles rainfall over, say, the last 2000 seasons, would you be concerned about constructing a confidence interval for the mean seasonal rainfall directly from the original data (with no transformation)?

E50. Display 9.47 shows a random sample of 7 seasons taken from the 128 seasons of Los Angeles rainfall in E49.

Display 9.47  Los Angeles rainfall in a random sample of 7 of the last 128 seasons.

a. Assuming you haven't seen the population in E49, are the conditions met for constructing a confidence interval for mean seasonal rainfall?

b. Regardless of your answer in part a, construct a 95% confidence interval.

c. The mean seasonal rainfall is 15.06 in. Did you capture the true mean in your confidence interval in part b?

d. Is the skewed shape of the population reflected in this small sample? If you take the log of each rainfall amount, are the conditions for inference better satisfied or worse? (What to do when you have a small sample isn't always clear. Fortunately, the \( t \)-procedure is fairly robust.)
E51. How should you measure the relationship between housing units (H) and the number of people living in them (P)—as persons per housing unit, or as housing units per person? The data in Display 9.48 give the population and number of housing units for a random sample of ten counties in Florida.

<table>
<thead>
<tr>
<th>County</th>
<th>Population (P) (thousands)</th>
<th>Housing Units (H) (thousands)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alachua</td>
<td>217.9</td>
<td>95.1</td>
</tr>
<tr>
<td>Collier</td>
<td>251.4</td>
<td>144.5</td>
</tr>
<tr>
<td>Duval</td>
<td>778.9</td>
<td>329.8</td>
</tr>
<tr>
<td>Hernando</td>
<td>430.8</td>
<td>62.7</td>
</tr>
<tr>
<td>Polk</td>
<td>483.9</td>
<td>226.4</td>
</tr>
<tr>
<td>Seminole</td>
<td>365.2</td>
<td>147.1</td>
</tr>
<tr>
<td>St. Lucie</td>
<td>192.7</td>
<td>91.3</td>
</tr>
<tr>
<td>Suwanee</td>
<td>34.8</td>
<td>15.7</td>
</tr>
<tr>
<td>Volusia</td>
<td>443.3</td>
<td>212.0</td>
</tr>
<tr>
<td>Walton</td>
<td>40.6</td>
<td>29.1</td>
</tr>
</tbody>
</table>

Display 9.48 Population and housing units in a sample of Florida counties. [Source: U.S. Census Bureau, Census 2000.]

a. Make a stemplot of the variable persons per housing unit (P/H) and another of housing units per person (H/P).

b. Construct and interpret a 95% confidence interval for the mean number of persons per housing unit.

c. Construct and interpret a 95% confidence interval for the mean number of housing units per person.

d. Turn the confidence interval in part c into one for the mean number of persons per housing unit by taking the reciprocals of the endpoints of the interval. How does this interval compare with the one in part b?

e. From parts a, b, and d, which variable do you think is better for measuring the relationship between housing units and their occupants?

E52. According to the popular press, some kinds of thinking (visual tasks) are “right-brained,” whereas others (verbal tasks) are “left-brained.” There is a fair amount of scientific support for this theory. For example, one experiment invented by L. R. Brooks involves two kinds of tasks. [Source: L. R. Brooks, “Spatial and Verbal Components of the Act of Recall,” Canadian Journal of Psychology 22 (1968): 349–68.]

The verbal task was to scan a sentence such as “The pencil is on the desk” and decide whether each word is or is not a noun. (The correct response is “No Yes No No No Yes.”) The visual task was to scan a block letter like the F shown here, starting at the arrow and moving clockwise, and decide whether each corner is an outside corner. (The correct response is “Yes Yes Yes No No Yes Yes No Yes.”)

Brooks also devised two ways to report, one verbal and one visual. To report verbally, you would simply say “Yes” or “No” out loud; to report visually, you would point in sequence to “Yes” or “No” on a piece of paper.

The theory predicted that when the task and the report were of the same kind, they would interfere with each other in memory and slow the response time. If visual and verbal tasks are handled independently, then a visual task with a verbal report or a verbal task with a visual report would be easier and so would take less time than a verbal task with a verbal report or a visual task with a visual report.

Display 9.49 (on the next page) shows the data from a version of the experiment run by a psychology lab at Mount Holyoke College.
<table>
<thead>
<tr>
<th>Subject</th>
<th>Visual Report Time (s)</th>
<th>Verbal Report Time (s)</th>
<th>Visual Report Time (s)</th>
<th>Verbal Report Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11.60</td>
<td>13.06</td>
<td>16.05</td>
<td>11.51</td>
</tr>
<tr>
<td>2</td>
<td>22.71</td>
<td>6.27</td>
<td>13.16</td>
<td>23.86</td>
</tr>
<tr>
<td>3</td>
<td>20.96</td>
<td>7.77</td>
<td>15.87</td>
<td>9.51</td>
</tr>
<tr>
<td>4</td>
<td>13.96</td>
<td>6.48</td>
<td>12.49</td>
<td>13.20</td>
</tr>
<tr>
<td>5</td>
<td>14.60</td>
<td>6.01</td>
<td>14.69</td>
<td>12.31</td>
</tr>
<tr>
<td>6</td>
<td>10.98</td>
<td>7.60</td>
<td>8.64</td>
<td>12.26</td>
</tr>
<tr>
<td>7</td>
<td>21.08</td>
<td>18.77</td>
<td>17.24</td>
<td>12.68</td>
</tr>
<tr>
<td>8</td>
<td>15.85</td>
<td>10.29</td>
<td>11.69</td>
<td>11.37</td>
</tr>
<tr>
<td>9</td>
<td>15.68</td>
<td>9.18</td>
<td>17.23</td>
<td>18.28</td>
</tr>
<tr>
<td>10</td>
<td>16.10</td>
<td>5.88</td>
<td>8.77</td>
<td>8.33</td>
</tr>
<tr>
<td>11</td>
<td>11.87</td>
<td>6.91</td>
<td>8.44</td>
<td>10.60</td>
</tr>
<tr>
<td>12</td>
<td>17.49</td>
<td>5.66</td>
<td>9.05</td>
<td>8.24</td>
</tr>
<tr>
<td>13</td>
<td>24.40</td>
<td>6.68</td>
<td>18.45</td>
<td>8.53</td>
</tr>
<tr>
<td>14</td>
<td>23.35</td>
<td>11.97</td>
<td>24.38</td>
<td>15.85</td>
</tr>
<tr>
<td>15</td>
<td>11.24</td>
<td>7.50</td>
<td>14.49</td>
<td>10.91</td>
</tr>
<tr>
<td>16</td>
<td>20.24</td>
<td>11.61</td>
<td>12.19</td>
<td>11.13</td>
</tr>
<tr>
<td>17</td>
<td>15.52</td>
<td>10.90</td>
<td>10.50</td>
<td>10.90</td>
</tr>
<tr>
<td>18</td>
<td>13.70</td>
<td>5.74</td>
<td>11.11</td>
<td>9.33</td>
</tr>
<tr>
<td>19</td>
<td>28.15</td>
<td>9.32</td>
<td>13.85</td>
<td>10.01</td>
</tr>
<tr>
<td>20</td>
<td>33.98</td>
<td>12.64</td>
<td>15.48</td>
<td>28.18</td>
</tr>
<tr>
<td>Mean</td>
<td>18.17</td>
<td>9.01</td>
<td>13.69</td>
<td>12.85</td>
</tr>
</tbody>
</table>

Display 9.49 Data on visual/verbal report times (in seconds) from the Mount Holyoke experiment.

a. The means of the four combinations of task and report are given in the last row of the table. Discuss the meaning of the pattern in light of the two predictions.

b. Boxplots showing the times for each of the four treatment groups are shown in Display 9.50. Are the distributions suitable for constructing a confidence interval for the mean, or would you recommend a transformation first? Does a reciprocal transformation make sense here? (To decide, compute the reciprocal of the visual task/visual report mean and interpret it.)

Display 9.50 Boxplots of the times of each of the four treatment groups.

c. Using the procedure you think best, estimate each of the four treatment means in 90% confidence intervals. Is there any evidence against the theory?

9.4 Inference for the Difference Between Two Means

Most scientific studies involve comparisons. For example, for purposes of studying pesticide levels in a river, does it make a difference whether you take water samples at mid-depth, near the bottom, or at the surface of the river? Do special exercises help babies learn to walk sooner? Inference about comparisons is more often used in scientific investigations than is inference about a single parameter. In fact, almost all experiments are comparative in nature. The study of inference for differences, then, is fundamental to statistical applications.
Warm-up: How Strong Is the Evidence?

Before you begin to develop formal methods of inference for differences between means, you can develop your intuition and your ability to make informal judgments based on comparing means by completing Activity 9.4a. It revisits the Wolf River data of P1 on page 573.

The Strength of Evidence

The Wolf River in Tennessee flows past an abandoned site once used by the pesticide industry for dumping wastes, including hexachlorobenzene (chlordane), aldrin, and dieldren. These highly toxic organic compounds can cause various cancers and birth defects. The standard method to test whether these poisons are present in a river is to take samples at six-tenths depth, that is, six-tenths of the way from the surface to the bottom. Unfortunately, there are good reasons to worry that six-tenths is the wrong depth. The organic compounds in question don’t have the same density as water, and their molecules tend to stick to particles of sediment. Both these facts suggest that you’d be likely to find higher concentrations near the bottom than near mid-depth.

1. Display 9.51 (on the next page) shows eight back-to-back stemplots. The first plot shows actual data. It compares concentrations of aldrin (in nanograms per liter) for 20 water samples taken from the Wolf River downstream from the dump site. Ten of the samples were taken at mid-depth, and ten were taken at the bottom. The other seven plots, numbered 1 through 7, show hypothetical data. Compare the plot of each hypothetical data set, 1 through 7, with the plot of the actual data and evaluate the strength of the evidence of a difference in concentration at the two depths. Is the evidence stronger in the actual data, stronger in the hypothetical data, or about the same for both? State your choice, and give the reason for your answer.

2. With everything else the same, is the evidence of a difference in concentration stronger when the difference in the means is smaller or larger? When the spread is smaller or larger? When the sample size is smaller or larger?
Most people can simply look at hypothetical data set 1 in Display 9.51 and correctly conclude that the mean concentration at the bottom probably isn't equal to the mean concentration at mid-depth. However, with the actual data and the other hypothetical data sets, it's not clear if there really is a difference. And to estimate the size of any difference, you need to compute a confidence interval.

A Confidence Interval for the Difference Between Two Means

A confidence interval for the difference between two means has the standard form you have seen before:

\[ \text{statistic} \pm (\text{critical value}) \cdot (\text{standard deviation of statistic}) \]

The procedure, given in the box, is also standard. Use this procedure when you want to estimate the size of the difference between the mean of one population and the mean of another population.

Confidence Interval for the Difference Between Two Means (Two-Sample t-Interval)

Check conditions.
You must check three conditions:

- The two samples were randomly and independently selected from two different populations. In the case of an experiment, the two treatments were randomly assigned to the available experimental units.
- The two samples look as if they came from normally distributed populations or the sample sizes are large enough that the sampling distributions of the sample means will be approximately normal. The 15/40 guideline (Section 9.3, pages 607–608) can be applied to each sample or treatment group, although it is a bit conservative.
- In the case of sample surveys, the population size should be at least ten times larger than the sample size for both samples.

Do computations.
A confidence interval for the difference between the means of two populations, $\mu_1 - \mu_2$, is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the respective means of the two samples, $s_1$ and $s_2$ are the standard deviations, and $n_1$ and $n_2$ are the sample sizes. It is best to use a calculator or statistics software to find this confidence interval because the value of $t^*$ depends on a complicated calculation.

Give interpretation in context and linked to computations.
For a 95% confidence interval, for example, you are 95% confident that if you knew the means of both populations, the difference between those means, $\mu_1 - \mu_2$, would lie in the confidence interval.

For an experiment, the interpretation is like this: If all experimental units could have been assigned each treatment, you are 95% confident that the difference between the means of the two treatment groups would lie in the confidence interval.

Of course, when you interpret a confidence interval, you do it in context, describing the two populations.

The standard error of the difference, $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, comes directly from the rules for random variables in the box in Section 6.1 on page 372. That is, if you have two independent random variables with variances $\left(\frac{s_1}{\sqrt{n_1}}\right)^2$, or $\left(\frac{s_1}{\sqrt{n_1}}\right)^2$, or $\left(\frac{s_2}{\sqrt{n_2}}\right)^2$, or $\left(\frac{s_2}{\sqrt{n_2}}\right)^2$, then the variance of the sampling distribution of their difference is found by adding the two variances, $\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$. Then take the square root to get the standard error, $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$. 
**Example: A Confidence Interval for the Aldrin Data**

Construct a 95% confidence interval for the difference between the mean bottom measurement of aldrin in the Wolf River and the mean mid-depth measurement. The data from the first stemplot in Display 9.51 are repeated in Display 9.52.

<table>
<thead>
<tr>
<th>Bottom</th>
<th>Mid-Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2 8</td>
</tr>
<tr>
<td>8</td>
<td>3 2 8</td>
</tr>
<tr>
<td>9 8</td>
<td>3 8 9</td>
</tr>
<tr>
<td>7 4 3</td>
<td>5 2 2</td>
</tr>
<tr>
<td>3 6</td>
<td>2 3 6</td>
</tr>
<tr>
<td>3 7</td>
<td></td>
</tr>
<tr>
<td>8 1</td>
<td></td>
</tr>
<tr>
<td>8 9</td>
<td></td>
</tr>
</tbody>
</table>

2 | 8 = 2.8 nanograms/liter

**Display 9.52** Actual aldrin concentrations, in nanograms per liter.

**Solution**

You will construct a confidence interval for the difference between two means (also called a two-sample \(t\)-interval). You have two samples from two populations, but you have no information about how randomly they were taken with respect to either their location in the river or time. Further, you don’t know if the samples were taken independently of one another. If, for example, a pair of bottom and mid-depth measurements were taken at the same time from the same spot, the measurements would not be independent (and you would use the techniques of the next section). Neither sample is skewed or has outliers, and each looks like it is reasonable to assume that it came from a normally distributed population.

Here are the summary statistics:

- **Bottom**: \( \bar{x}_1 = 6.04 \), \( s_1 = 1.579 \), \( n_1 = 10 \)
- **Mid-depth**: \( \bar{x}_2 = 5.05 \), \( s_2 = 1.104 \), \( n_2 = 10 \)

Using a calculator (which does not give \( t^* \)), you get a 95% confidence interval of

\[
(\bar{x}_1 - \bar{x}_2) \pm t^* \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (6.04 - 5.05) \pm t^* \cdot \sqrt{\frac{(1.579)^2}{10} + \frac{(1.104)^2}{10}}
\]

\[
= 0.99 \pm 1.2909
\]

or \((-0.3009, 2.2809)\), where \(df = 16.10\).

You are 95% confident that the true difference between the mean aldrin levels in the two depths in the Wolf River, \( \mu_{\text{bottom}} - \mu_{\text{mid-depth}} \), is in the interval \((-0.3009, 2.2809)\). This interval overlaps 0 slightly, so there is insufficient evidence to say that the true mean of the bottom measurements differs from the true mean of the mid-depth measurements (but it is close). However, just as for any confidence interval, unless the conditions are satisfied, there is no automatic guarantee that the capture rate will be equal to the advertised confidence level. Thus, the conclusion must be limited to the population of potential measurements from which these samples were taken.
More on Independence. The aldrin samples might not have been taken independently. That would be the case if two measurements were taken at the same time at the same spot in the river, one at mid-depth and one at the bottom, and this procedure was repeated at nine other locations. When the measurements of the two samples are taken in pairs (such as measurements from husband/wife or left foot/right foot), they are not independent, and you should use the methods of the next section.

More on Sample Sizes. The 15/40 guideline for inference for the difference of two means is a bit conservative. In other words, the minimum sample sizes of 15 and 40 are higher than they need to be, in most cases. The reason is that, with skewed populations, the sampling distribution of the difference of two means tends to be more symmetric than the two separate sampling distributions of the sample mean. Subtracting two sample means brings in the tails. The net effect is that the sampling distribution of the difference of two means will look approximately normal for smaller sample sizes than would be the case for the means themselves.

More on Degrees of Freedom. In the aldrin example, the calculator gave a value for degrees of freedom, 16.10, that isn't a whole number. What's that all about? The matter is fairly complicated, as you might have guessed, and that's why it is better to let the calculator find the confidence interval. The basic idea is that, unlike in the one-sample case, the sampling distribution of the statistic for the difference of two samples doesn't have a \( t \)-distribution. The exact distribution isn't even known. However, it is known that the distribution is reasonably close to a \( t \)-distribution if the right number of degrees of freedom is used. If you are interested in the formula your calculator uses to compute these degrees of freedom, see D14 on page 624.

In Which Order Do You Subtract? Usually, it doesn’t matter whether you compute a confidence interval for \( \mu_1 - \mu_2 \) or for \( \mu_2 - \mu_1 \), so, if it seems reasonable, subtract in the order that makes the difference positive and thus easier to interpret. Just be sure to keep the SDs and sample sizes with their respective means, and be sure your interpretation reflects the order of subtraction you used.

Example: Walking Babies 1

We’ll use the same scenario and experimental data as in E2 on page 576. The effects of special exercises to speed up infant walking were isolated using three different control groups. Display 9.53 shows the data again.

<table>
<thead>
<tr>
<th>Group</th>
<th>Age (in months) at First Unaided Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Special exercises</td>
<td>9 9.5 9.75 10 13 9.5</td>
</tr>
<tr>
<td>Exercise control</td>
<td>11 10 10 11.75 10.5 15</td>
</tr>
<tr>
<td>Weekly report control</td>
<td>11 12 9 11.5 13.25 13</td>
</tr>
<tr>
<td>Final report control</td>
<td>13.25 11.5 12 13.5 11.5</td>
</tr>
</tbody>
</table>


Use these data to find a 95% confidence interval estimate of the difference between mean walking times for the special exercises group and the exercise control group if all babies could have had each treatment.
**Solution**

Two treatments were randomly assigned to the babies. You should plot the data from the two groups to see if there is any reason to doubt that they resemble samples from normally distributed populations. Display 9.54 shows that there is some doubt, because each group has one relatively large value. Further, we were told that some babies are 18 months old before they walk, so the population of ages must be somewhat skewed right. However, the sampling distribution of the difference of the means tends to be more symmetric than that of either group mean by itself, so a confidence interval for the difference of two means should work for this amount of skewness.

![Dot plots of two samples](image)

**Display 9.54** Dot plots of two samples.

**Do computations.**

Special exercises: \( \bar{x}_1 = 10.125 \quad n_1 = 6 \quad s_1 = 1.447 \)

Exercise control: \( \bar{x}_2 = 11.375 \quad n_2 = 6 \quad s_2 = 1.896 \)

The two-sample \( t \)-interval function of a calculator gives \( df \approx 9.35 \) and the confidence interval \((-3.44, 0.94)\). [See Calculator Note 9F to learn how to calculate a two-sample \( t \)-interval. You can start with the actual data or summary statistics.]

![Calculator output](image)

You are 95% confident that if all the babies could have been in the special exercises group and all the babies could have been in the exercise control group, the difference in the mean age at which they would learn to walk is in the interval from \(-3.44\) months to 0.94 month. Because 0 is in the confidence interval, you have no evidence that there would be any difference if you were able to give each treatment to all the babies. However, it is important to note that any conclusion is subject to doubt about the appropriateness of this procedure due to skewness of the population of potential measurements, so it would be good to confirm this finding with more data.

Display 9.55 shows printouts for the two-sample \( t \)-interval for the walking babies experiment from three commonly used statistical software packages.
Data Desk

2-Sample t-Interval for $\mu_1 - \mu_2$
No Selector
Individual Confidence 95.00%
Bounds: Lower Bound < $\mu_1 - \mu_2$ < Upper Bound
With 95.00% Confidence, $-3.4399682 < \mu(\text{Special}) - \mu(\text{Exercise}) < 0.9399682$

Fathom

Estimate of Walking Babies

<table>
<thead>
<tr>
<th>First attribute (numeric): Special_Exercise</th>
<th>Difference of Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Second attribute (numeric or categorical): Exercise_Control</td>
<td></td>
</tr>
<tr>
<td>Interval estimate for the population mean of Special_Exercise minus that of Exercise_Control</td>
<td></td>
</tr>
<tr>
<td>Count:</td>
<td></td>
</tr>
<tr>
<td>Special_Exercise</td>
<td>Exercise_Control</td>
</tr>
<tr>
<td>Mean:</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Std dev:</td>
<td></td>
</tr>
<tr>
<td>1.44698</td>
<td>1.89572</td>
</tr>
<tr>
<td>Std error:</td>
<td></td>
</tr>
<tr>
<td>0.890727</td>
<td>0.773924</td>
</tr>
<tr>
<td>Confidence level: 95.0</td>
<td></td>
</tr>
<tr>
<td>Using unpoold variances</td>
<td></td>
</tr>
<tr>
<td>Estimate: $-1.25 +/- 2.18997$</td>
<td></td>
</tr>
<tr>
<td>Range: $-3.43997$ to $0.939967$</td>
<td></td>
</tr>
</tbody>
</table>

Minitab

TWOSAMPLE T FOR Special VS Exercise

<table>
<thead>
<tr>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SEMean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Special</td>
<td>6</td>
<td>10.12</td>
<td>1.45</td>
</tr>
<tr>
<td>Exercise</td>
<td>6</td>
<td>11.38</td>
<td>1.90</td>
</tr>
</tbody>
</table>

95 PCT CI FOR MU Special - MU Exercise: (-3.45, 0.95)

TTTEST MU Special = MU Exercise (VS NE): T=-1.28 P=0.23 DF= 9

Display 9.55 Printouts of estimation from three statistical software packages.

A Confidence Interval for the Difference Between Two Means

D13. Here are summary statistics for the hypothetical data sets of Activity 9.4a. Assume samples are random, observations are independent, and populations are normally distributed.

i. Construct a 95% confidence interval for the difference between the two means, bottom — mid-depth, for each data set.

ii. Are the results consistent with the intuitive judgments you made about the strength of the evidence in Activity 9.4a?

a. Hypothetical Data Set 1

Bottom: $\bar{x}_1 = 7.04$ \hspace{1cm} $n_1 = 10$ \hspace{1cm} $s_1 = 1.6$

Mid-depth: $\bar{x}_2 = 5.05$ \hspace{1cm} $n_2 = 10$ \hspace{1cm} $s_2 = 1.1$
b. Hypothetical Data Set 2

Bottom: $\bar{x}_1 = 6.04$  $n_1 = 10$  $s_1 = 1.3$
Mid-depth: $\bar{x}_2 = 5.05$  $n_2 = 10$  $s_2 = 0.8$

c. Hypothetical Data Set 3

Bottom: $\bar{x}_1 = 6.04$  $n_1 = 10$  $s_1 = 1.9$
Mid-depth: $\bar{x}_2 = 5.75$  $n_2 = 10$  $s_2 = 1.8$

d. Hypothetical Data Set 4

Bottom: $\bar{x}_1 = 6.04$  $n_1 = 20$  $s_1 = 1.5$
Mid-depth: $\bar{x}_2 = 5.05$  $n_2 = 20$  $s_2 = 1.1$

e. Hypothetical Data Set 5

Bottom: $\bar{x}_1 = 5.97$  $n_1 = 7$  $s_1 = 1.7$
Mid-depth: $\bar{x}_2 = 5.07$  $n_2 = 6$  $s_2 = 1.3$

f. Hypothetical Data Set 6

Bottom: $\bar{x}_1 = 7.04$  $n_1 = 10$  $s_1 = 1.6$
Mid-depth: $\bar{x}_2 = 6.05$  $n_2 = 10$  $s_2 = 1.1$

g. Hypothetical Data Set 7

Bottom: $\bar{x}_1 = 6.04$  $n_1 = 10$  $s_1 = 1.6$
Mid-depth: $\bar{x}_2 = 6.05$  $n_2 = 10$  $s_2 = 1.1$

D14. Your calculator or statistics software uses this formula to find $df$ when doing a two-sample $t$-procedure. You might be curious as to why we need this complicated rule to calculate degrees of freedom. The simple answer is that using $t$ in place of $z$ does not provide quite the right adjustment in the two-sample case, as it does in the one-sample case, unless we make this additional adjustment to the degrees of freedom. (The whole theoretical story is complicated.)

\[
\frac{(\sigma_{\bar{x}_1-\bar{x}_2})^2}{df} = \frac{(\sigma_{\bar{x}_1})^2}{df_1} + \frac{(\sigma_{\bar{x}_2})^2}{df_2}
\]

\[
\frac{\left(\frac{s^2}{n_1} + \frac{s^2}{n_2}\right)^2}{df} = \frac{\left(\frac{s^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s^2}{n_2}\right)^2}{n_2 - 1}
\]

a. Verify the value of $df$ given in the aldrin and walking babies examples on pages 620 and 621.

b. If $n_1 = n_2$, derive a simplified version of the formula for $df$.

c. If $n_1 = n_2$ and, in addition, $s_1 = s_2$, derive an even simpler rule for $df$.

**A Significance Test for the Difference of Two Means**

As you know, the other way to approach an inference problem is through significance testing. Here’s a summary of significance testing for the difference between two means.
Components of a Test for the Difference Between Two Means (the Two-Sample t-Test)

1. **Name the test and check conditions.** For a test of significance of the difference between two means, the methods of this section require the same conditions as those for a confidence interval:
   - Samples have been randomly and independently selected from two different populations. In the case of an experiment, the two treatments were randomly assigned to the available experimental units.
   - The two samples look as if they come from normally distributed populations or the sample sizes are large enough that the sampling distributions of the sample mean will be approximately normal. The 15/40 guideline can be applied to each sample, although it is a bit conservative. As noted in the conditions for confidence intervals, you can get by with smaller sample sizes when taking a difference.
   - In the case of sample surveys, the population size should be at least ten times larger than the sample size for both samples.

2. **State your hypotheses.** The null hypothesis is, ordinarily, that the two population means are equal. In symbols, $H_0: \mu_1 = \mu_2$ or, in terms of the difference between the means, $H_0: \mu_1 - \mu_2 = 0$. Here $\mu_1$ is the mean of the first population and $\mu_2$ is the mean of the second. There are three forms of the alternative or research hypothesis:
   - $H_a: \mu_1 \neq \mu_2$ or $H_a: \mu_1 - \mu_2 \neq 0$
   - $H_a: \mu_1 < \mu_2$ or $H_a: \mu_1 - \mu_2 < 0$
   - $H_a: \mu_1 > \mu_2$ or $H_a: \mu_1 - \mu_2 > 0$

3. **Compute the test statistic, find the P-value, and draw a sketch.** Compute the difference between the sample means (because the hypothesized mean difference is zero), measured in estimated standard errors:

\[
t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}
\]

Use the two-sample t-test function of your calculator or statistics software to get the P-value.

4. **Write your conclusion, linked to your computations and in the context of the problem.** If you are using fixed-level testing, reject the null hypothesis if your P-value is less than the level of significance, $\alpha$. If the P-value is greater than or equal to $\alpha$, do not reject the null hypothesis. (If you are not given a value of $\alpha$, you can assume that $\alpha$ is 0.05.) Write a conclusion that relates to the situation and includes an interpretation of your P-value.
Example: A Significance Test for the Aldrin Data

You’ve been given the responsibility to analyze the Wolf River data in Display 9.51 on page 618 to test whether the true mean aldrin concentrations at the bottom and at mid-depth might differ. That is, you want to set up a test of significance for the difference between two population means based on data from independent random samples. Use $\alpha = 0.10$.

Solution

Check conditions.

Conditions were checked in the example on page 620, and the conclusion here is subject to those concerns.

State your hypotheses.

In terms of the difference between two means, your null hypothesis is

$$H_0: \mu_{\text{bottom}} = \mu_{\text{mid-depth}} \quad \text{or} \quad H_0: \mu_{\text{bottom}} - \mu_{\text{mid-depth}} = 0$$

where $\mu_{\text{bottom}}$ is the mean aldrin concentration at the bottom of the Wolf River and $\mu_{\text{mid-depth}}$ is the mean concentration at six-tenths depth.

You are looking for a difference in either direction, so the alternative hypothesis is two-sided:

$$H_a: \mu_{\text{bottom}} \neq \mu_{\text{mid-depth}} \quad \text{or} \quad H_a: \mu_{\text{bottom}} - \mu_{\text{mid-depth}} \neq 0$$

Here are the summary statistics for the aldrin concentrations:

- Bottom: $\bar{x}_1 = 6.04$, $n_1 = 10$, $s_1 = 1.579$
- Mid-depth: $\bar{x}_2 = 5.05$, $n_2 = 10$, $s_2 = 1.104$

The value of the test statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(6.04 - 5.05) - 0}{\sqrt{\frac{1.579^2}{10} + \frac{1.104^2}{10}}} \approx 1.625$$

From a calculator, you get an approximate $df$ of 16.10 and a $P$-value of 0.1236. [See Calculator Note 9G.]

Give conclusion in context.

Conclude that, because the $P$-value for a two-sided test is greater than $\alpha = 0.10$, you do not reject the null hypothesis. There is insufficient evidence to claim that the mean aldrin concentration at the bottom of the Wolf River is
different from the mid-depth concentration. In other words, although it appears from the stem-and-leaf plot that aldrin concentrations are greater near the bottom, with these sample sizes the difference is not large enough to rule out chance variation as a possible explanation.

Display 9.56 shows computer printouts from three commonly used statistical software packages for the analysis in this example.

**Data Desk**

2-Sample t-Test of $\mu_1 - \mu_2$
No Selector
Individual Alpha Level 0.10
H0: $\mu_1 - \mu_2 = 0$ Ha: $\mu_1 - \mu_2 \neq 0$
bottom - middepth
Test H0: $\mu(\text{bottom}) - \mu(\text{middepth}) = 0$ vs Ha: $\mu(\text{bottom}) - \mu(\text{middepth}) \neq 0$
Difference Between Means = 0.99000000 t-Statistic = 1.625 w/16 df
Fail to reject H0 at Alpha = 0.10
p = 0.1236

**Fathom**

Test of Wolf River

<table>
<thead>
<tr>
<th>First attribute (numeric): Bottom</th>
<th>Second attribute (numeric or categorical): Mid_Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ho: Population mean of Bottom equals that of Mid_Depth</td>
<td></td>
</tr>
<tr>
<td>Ha: Population mean of Bottom is not equal to that of Mid_Depth</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bottom</th>
<th>Mid_Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count:</td>
<td>10</td>
</tr>
<tr>
<td>Mean:</td>
<td>6.04</td>
</tr>
<tr>
<td>Std dev:</td>
<td>1.57917</td>
</tr>
<tr>
<td>Std error:</td>
<td>0.499377</td>
</tr>
</tbody>
</table>

Using *unpooled variances*

Student's t: 1.625
DF: 16.0994
P-value: 0.12

**Minitab**

TWOSAMPLE T FOR Bottom VS MidDepth

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SEMean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottom</td>
<td>10</td>
<td>6.04</td>
<td>1.58</td>
<td>0.50</td>
</tr>
<tr>
<td>Middepth</td>
<td>10</td>
<td>5.05</td>
<td>1.10</td>
<td>0.35</td>
</tr>
</tbody>
</table>

TTEST MU Bottom = MU Middepth (VS NE): T=1.62 P=0.12 DF=16

**Display 9.56** Printouts of significance tests from three statistical software packages.

For the aldrin data, it makes sense to use a two-sided alternative. (Researchers wanted to know whether samples taken at mid-depth would give essentially the same results as samples taken near the bottom.) For the walking babies
experiment, a one-sided alternative makes more sense because the researchers wanted to test whether the special exercises helped babies walk sooner. They weren’t interested in the possibility that the special exercises would slow down the babies.

**Example: Walking Babies 2**

Test the null hypothesis that the mean age at first unaided steps is the same for the special exercises and exercise control groups against the alternative hypothesis that the mean age is smaller for the special exercises group. Use \( \alpha = 0.05 \).

**Solution**

The conditions were checked in the example “Walking Babies 1” on page 621.

The null hypothesis is

\[
H_0: \mu_{\text{special ex}} = \mu_{\text{control ex}} \quad \text{or, equivalently,} \quad H_0: \mu_{\text{special ex}} - \mu_{\text{control ex}} = 0
\]

where \( \mu_{\text{special ex}} \) is the mean age that the babies in the experiment would first walk if they all could have been given the special exercises and \( \mu_{\text{control ex}} \) is the mean age that the babies would first walk if all babies in the experiment could have received the exercise control treatment. (You can also use the symbols \( \mu_1 \) and \( \mu_2 \) as long as you define the symbols.)

For this one-sided test, the alternative hypothesis is

\[
H_a: \mu_{\text{special ex}} < \mu_{\text{control ex}} \quad \text{or, equivalently,} \quad H_a: \mu_{\text{special ex}} - \mu_{\text{control ex}} < 0
\]

Here are the summary statistics:

- Special exercises: \( \bar{x}_1 = 10.125 \quad n_1 = 6 \quad s_1 = 1.447 \)
- Exercise control: \( \bar{x}_2 = 11.375 \quad n_2 = 6 \quad s_2 = 1.896 \)

The test statistic is

\[
t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(10.125 - 11.375) - 0}{\sqrt{\frac{1.447^2}{6} + \frac{1.896^2}{6}}} = -1.284
\]

From a calculator, \( df \) is 9.35 and the \( P \)-value is 0.115. Because the \( P \)-value, 0.115, is greater than \( \alpha = 0.05 \), you do not reject the null hypothesis. The evidence isn’t convincing that babies who are given the special exercises walk at an earlier age than babies who are given the control exercise. In other words, if the researchers had been able to give both the special exercises and the control exercises to all the babies in the experiment, they would have been reasonably likely to get groups as different as these even if the special exercises made no difference.
Increasing Power in the Two-Sample t-Test

As always, the best way to get more power to reject a false null hypothesis is to increase the sample sizes. However, sometimes you have the resources to get, say, only 40 measurements total between the two samples. How should you divide them up—20 in each sample—or some other way?

Getting the Most Power Out of Your Two-Sample t-Test

The best way to get more power is to have larger sample sizes.

If you have reason to believe that the population standard deviations are about equal, make the sample sizes equal.

If you have reason to believe that one population’s standard deviation is larger than the other’s, allocate your resources so that you take a larger sample from the population with the larger standard deviation. (Choose the sample sizes to be proportional to the estimated standard deviations.)

Follow these same rules to get the smallest margin of error for a confidence interval.

Why Not Two Separate Confidence Intervals? Significance tests for the difference of two means take a little getting used to. It is natural to ask why you can’t simply compute two separate confidence intervals, one for $\mu_1$ and one for $\mu_2$, and check to see if they overlap. This method will tell you if there are any values that are plausible means for both populations. If so, then you wouldn’t reject the null hypothesis that the difference in the means is 0. The difficulty is that the method is too conservative, meaning that you won’t reject a false null hypothesis often enough. In other words, you have sacrificed power.

However, you can use this rule: If you construct two separate confidence intervals for the means of two populations (at confidence level $1 - \alpha$) and they don’t overlap, you are safe in rejecting the null hypothesis that the means are equal at significance level $\alpha$. If the intervals overlap, you can come to no conclusion.

A Special Case: Pooling When $\sigma_1 = \sigma_2$

Suppose you are taking two independent samples from the same population for the purpose of comparing means. This happens, for example, when you randomly divide available experimental units into two groups for the purpose of comparing two treatments. If the true treatment means really do not differ (the usual null hypothesis), then the true variances of the sample measurements should not differ either. You then have two sample variances (that probably will differ) to estimate the single population variance. One way to combine these two estimates is simply to average the sample variances and use this average to estimate the population variance. (This average should be a weighted average if the sample sizes are not equal.) This process, called pooling, gives you another way to calculate a confidence interval or a test statistic based on $t$.

Your calculator and statistics software give you the choice of “pooled” or “unpooled” when doing two-sample $t$-procedures. “Pooled” should be used only when you have a good reason to believe that the population standard deviations
are equal. Even if you know this to be true, the two-sample (unpooled) procedure discussed in this chapter works almost as well as the pooled procedure, especially if the sample sizes are equal. The only situation in which the pooled procedure has a definite advantage is when the populations have equal standard deviations but your sample sizes are unequal. Therefore, we won’t cover the pooled procedure in this text. Unless you encounter a problem that specifically tells you to assume that \( \sigma_1 \) and \( \sigma_2 \) are equal, choose the unpooled procedure.

You should know, however, that there are extensions of the two-sample procedures that allow comparisons among more than two means. The most common of these procedures (analysis of variance, or ANOVA) is a generalization of the pooled procedure for two samples, so it is good to know when pooling works and when it doesn’t.

**Increasing Power in the Two-Sample t-Test**

D15. *To pool or not to pool?* Here are three different situations for two independent samples.

I. \( n_1 = 5, n_2 = 25, \sigma_1 = 10, \sigma_2 = 10 \)

II. \( n_1 = 10, n_2 = 10, \sigma_1 = 10, \sigma_2 = 10 \)

III. \( n_1 = 5, n_2 = 25, \sigma_1 = 10, \sigma_2 = 1 \)

For one of these situations, pooling is wrong. For a second situation, pooling, though not wrong, is not likely to offer an advantage over the unpooled approach. Finally, in the third situation, pooling not only is appropriate but also will likely give narrower intervals than the unpooled approach. Which is which?

**Summary 9.4: Inference for the Difference Between Two Means**

In this section, you learned how to construct a confidence interval for \( \mu_1 - \mu_2 \) and to perform a significance test of \( \mu_1 = \mu_2 \). Whether you are doing a confidence interval or a significance test, the conditions to check are the same:

- Samples have been randomly and independently selected from two different populations (or two treatments were randomly assigned to the available experimental units).
- The two samples look as if they came from normally distributed populations or the sample sizes are large enough. (The 15/40 guideline can be applied to each sample, although it is a bit conservative.)
- In the case of sample surveys, the population size should be at least ten times larger than the sample size for both samples.

A confidence interval for the difference between the means of two populations, \( \mu_1 - \mu_2 \), has the form

\[
(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}
\]
The test statistic for a test of significance is

\[ t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \]

This statistic is called \( t \), but it doesn't have exactly a \( t \)-distribution. Fortunately, it is reasonably accurate to proceed using the \( t \)-distribution with \( df \) approximated by a rather complicated rule. A calculator or statistics software is the most accurate method of computing a confidence interval and of finding a \( P \)-value. Always select the unpooled \( t \)-procedure unless you are asked to do otherwise.

### Practice

#### A Confidence Interval for the Difference Between Two Means

P27. As you read in E1 on page 575, some students recruited 30 volunteers to attempt to walk the length of a football field while blindfolded. Each volunteer began at the middle of one goal line and was asked to walk to the opposite goal line, a distance of 100 yards. The dot plot and summary statistics in Display 9.57 show the distance at which the volunteer crossed a sideline of the field and whether the volunteer was left-handed or right-handed.

a. Assuming that these volunteers can be considered independent random samples from the populations of left-handed volunteers and right-handed volunteers, are the conditions met for a confidence interval for the difference of two means?

b. Regardless of your answer to part a, construct a 95% confidence interval, using your calculator or statistics software.

c. You are 95% confident that something is in the interval you constructed in part b. Describe exactly what that something is.

d. Do you have statistically significant evidence that left- and right-handed volunteers differ in the mean number of yards they can walk before crossing a sideline?

P28. The data in Display 9.58 (on the next page) are from a survey done in an introductory statistics class during the first week of a term (see P6 on page 574). You can assume that this class is a random sample taken from all students in this course. The data are the number of hours of study per week, classified by gender.

Estimate the difference in mean study hours per week for all the males and females taking this course, with confidence level 0.90. Interpret this interval.
Gender  N  Mean  Median  StDev
Study Hours F 46  10.93  10.00  6.22
        M 15  8.20  7.00  5.94

Stem-and-leaf of Study
Gender = Female N = 46
Leaf Unit = 1.0
5  0  22233
10  0  55555
16  0  666777
19  0  888
(b) 1  0000011
19  1  22233
14  1  555555
  8  1
8  1  88
6  2  0000
  2  3
  1  2
  1  2
  1  3

Display 9.58  Two stemplots of study hours per week, classified by gender.

A Significance Test for the Difference of Two Means

P29. Refer to P28. Test to see if there is evidence of a statistically significant difference in mean weekly study hours for males and females at the 5% level of significance.

P30. Refer to the walking babies data in Display 9.53 on page 621.
   a. Test to see if the special exercises group produces a statistically significant decrease in mean walking age over the weekly report group. Use $\alpha = 0.05$.
   b. Test to see if the exercise control group produces a statistically significant decrease in mean walking age over the weekly report group. Use $\alpha = 0.05$.
   c. State a conclusion about comparisons among the three groups.

Exercises

E53. Kelly randomly assigned eight golden hamsters to be raised in long days or short days. She then measured the concentrations of an enzyme in their brains. (Refer to page 244 for more about Kelly’s hamster experiment.) The resulting measurements of enzyme concentrations (in milligrams per 100 milliliters) for the eight hamsters (shown in Display 9.60) were

<table>
<thead>
<tr>
<th>Bottom</th>
<th>Mid-Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2.8</td>
</tr>
<tr>
<td>8</td>
<td>3 2.8</td>
</tr>
<tr>
<td>9 8</td>
<td>4 3.8 9</td>
</tr>
<tr>
<td>7 4 3</td>
<td>5 2.2</td>
</tr>
<tr>
<td>3 6</td>
<td>2.3 6</td>
</tr>
<tr>
<td>3 7</td>
<td></td>
</tr>
<tr>
<td>8 1</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>

| 2 | 8 = 2.8 nanograms/liter |

Display 9.59  Actual aldrin concentrations.

P31. The actual aldrin data appear in Display 9.59. Your boss sends you out to get ten more measurements. Would you suggest getting five additional measurements at each depth, or something different? Explain.
9.4 Inference for the Difference Between Two Means

Display 9.60 A dot plot of Kelly’s hamster data.

a. Are the conditions met for inference about the difference of two means?

b. Regardless of your answer to part a, construct a 95% confidence interval.

c. You are 95% confident that something is in the interval you constructed in part b. Describe exactly what that something is.

d. Does Kelly have statistically significant evidence to back her claim that the observed difference in enzyme concentrations between the two groups of hamsters is due to the difference in the hours of daylight?

E54. Suppose Kelly’s means and the shapes of the distributions are the same as in E53 but the enzyme concentrations are more variable, as shown in Display 9.61.

Short days: 9.500 8.625 27.275 10.225

Long days: 4.625 12.375 11.900 6.800

Display 9.61 Dot plots of altered hamster data.

a. With the altered data, are the conditions met for inference about the difference of two means?

b. Regardless of your answer to part a, construct and interpret a 95% confidence interval.

c. If the values had been this variable, would Kelly have had statistically significant evidence to back her claim that the observed difference in enzyme concentrations between the two groups of hamsters is due to the difference in the hours of daylight?

E55. The inflammation caused by osteoarthritis of the knee can be very painful and can inhibit movement. Leech saliva contains anti-inflammatory substances. To study the therapeutic effect of attaching four to six leeches to the knee for about 70 minutes, 51 volunteers were randomly assigned to receive either the leech treatment or a topical gel, diclofenac. [Source: Andreas Michalsen et al., “Effectiveness of Leech Therapy in Osteoarthritis of the Knee: A Randomized, Controlled Trial,” Annals of Internal Medicine 139, no. 9 (November 4, 2003): 724–30.]

a. This summary table gives the results of a pretreatment measure of the amount of pain reported by the two groups, before beginning therapy. A higher score means more pain. The researchers hoped that the randomization would result in two comparable groups with respect to this variable. Construct and interpret a 95% confidence interval for the difference in means. Is there statistically significant evidence that the randomization failed to yield groups with comparable means?

Leech: $\bar{x}_1 = 53.0 \quad s_1 = 13.7 \quad n_1 = 24$

Topical gel: $\bar{x}_2 = 51.5 \quad s_2 = 16.8 \quad n_2 = 27$

b. Because a high body mass can stress the knee, the researchers hoped that the randomization would result in two comparable groups with respect to this variable as well. This summary table gives the body mass index of the two treatment groups, before beginning therapy. Construct and interpret a 95% confidence interval for the difference in means. Is there statistically significant evidence that the randomization failed to yield groups with comparable means?

Leech: $\bar{x}_1 = 27.6 \quad s_1 = 3.7 \quad n_1 = 24$

Topical gel: $\bar{x}_2 = 27.1 \quad s_2 = 3.7 \quad n_2 = 27$
E56. Refer to E55. So far the researchers have been lucky, and the two treatment groups have comparable means. However, the researchers compared the two treatment groups on 12 initial variables. Another variable was an initial measure of stiffness.

a. This summary table gives summary statistics for the stiffness scores of the two groups, before beginning therapy. A higher score means more stiffness. Construct and interpret a 95% confidence interval for the difference in means. Is there statistically significant evidence that the randomization failed to yield groups with comparable means?

Leech: \( \bar{x}_1 = 63.3 \quad s_1 = 19.0 \quad n_1 = 24 \)

Topical gel: \( \bar{x}_2 = 48.6 \quad s_2 = 22.2 \quad n_2 = 27 \)

b. If you can consider the initial variables to be independent, how many of the 12 variables would you expect to show a statistically significant difference between two randomly assigned groups?

c. If you can consider the initial variables to be independent, what is the probability that none of the 12 variables shows a statistically significant difference between two randomly assigned groups?

d. At the end of the experiment, the leech treatment was shown to be significantly better than the topical gel. Does your result in part a tend to invalidate this conclusion?

E57. In E19 and E20 on page 598, students were given pieces of paper with five vertical or horizontal line segments and asked to mark the midpoint of each segment. Then, using a ruler, each student measured how far each mark was from the real midpoint of each segment. He or she then averaged the five errors. (We aren’t told if the assignments of the two treatments were random, but assume they were because this experiment was done in a statistics class.)

The results are given in Display 9.62.

### Display 9.62 Average error in locating midpoints of horizontal and vertical line segments.

<table>
<thead>
<tr>
<th>Vertical (positive means too high)</th>
<th>Horizontal (positive means too far right)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.6</td>
<td>−0.9</td>
</tr>
<tr>
<td>1.2</td>
<td>−0.4</td>
</tr>
<tr>
<td>1.1</td>
<td>−0.4</td>
</tr>
<tr>
<td>1.1</td>
<td>−0.2</td>
</tr>
<tr>
<td>1.0</td>
<td>−0.2</td>
</tr>
<tr>
<td>0.8</td>
<td>0.0</td>
</tr>
<tr>
<td>0.4</td>
<td>0.0</td>
</tr>
<tr>
<td>0.4</td>
<td>0.0</td>
</tr>
<tr>
<td>0.4</td>
<td>0.2</td>
</tr>
<tr>
<td>0.2</td>
<td>0.4</td>
</tr>
<tr>
<td>0.2</td>
<td>0.4</td>
</tr>
<tr>
<td>0.0</td>
<td>0.4</td>
</tr>
<tr>
<td>0.0</td>
<td>0.8</td>
</tr>
<tr>
<td>−0.4</td>
<td></td>
</tr>
</tbody>
</table>

a. Before beginning this experiment, the instructor believed that students tend to make larger errors (in absolute value) when the line segments are vertical than when the line segments are horizontal. Make the necessary changes to the data in the table so you can test the instructor’s belief.

b. Conduct a significance test of the instructor’s claim, showing all four steps.

E58. It has been speculated that the mean amount of calcium in the blood is higher in women than in men. A retrospective review of the medical charts of subjects tested in a certain city produced the summary statistics on calcium (in millimoles per liter) shown in Display 9.63.

a. Discuss whether the conditions are met for doing a two-sample \( t \)-test of the difference in the means.
b. Find and interpret the P-value for this test.

<table>
<thead>
<tr>
<th>Calcium</th>
<th>Sex</th>
<th>Row Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>F</td>
</tr>
<tr>
<td>2.3181319</td>
<td>2.3969048</td>
<td>2.3559429</td>
</tr>
<tr>
<td>91</td>
<td>84</td>
<td>175</td>
</tr>
<tr>
<td>0.12172749</td>
<td>0.14049805</td>
<td>0.13852122</td>
</tr>
<tr>
<td>0.012760508</td>
<td>0.015329954</td>
<td>0.010320034</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Display 9.63 Calcium in the blood of women and men, in millimoles per liter. [Source: JSE Data Archive, www.amstat.org.]

E59. In warm and humid parts of the world, a constant battle is waged against termites. Scientists have discovered that certain tree resins are deadly to termites, and thus the trees producing these resins become a valuable crop. In one experiment typical of the type used to test the protective power of a resin, two doses of resin (5 mg and 10 mg) were dissolved in a solvent and placed on filter paper. Eight dishes were prepared with filter paper at dose level 5 mg and eight with filter paper at dose level 10 mg. Twenty-five termites were then placed in each dish to feed on the filter paper. At the end of 15 days, the number of surviving termites was counted. The results are shown in Display 9.64. In parts a–c, you’ll determine if there is a statistically significant difference in the mean number of survivors for the two doses.

a. Are the conditions for a two-sample t-test met here?

b. Regardless of your answer to part a, carry out an appropriate test. Carefully write your conclusion in the context of the study.

c. What are your main concerns about the conclusion you reached in part b? What advice would you give the experimenter?
E60. The fact that your food usually tastes good is no accident. Food manufacturers regularly check taste and texture by recruiting taste-test panels to measure palatability.

A standard method is to form a panel with 50 persons—25 men and 25 women—to do the tasting. In one such experiment, simplified here, coarse versus fine texture was compared. Panel members were assigned randomly to the treatment groups as they were recruited. There were 16 panels of 50 consumers each. The variables in Display 9.65 are these:

- **Total palatability score for the panel of 50:** A higher total score means the food was rated more palatable by the panel.
- **Texture:** 0 (coarse), 1 (fine)

<table>
<thead>
<tr>
<th>Score</th>
<th>Texture</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>0</td>
</tr>
<tr>
<td>39</td>
<td>0</td>
</tr>
<tr>
<td>77</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>104</td>
<td>1</td>
</tr>
<tr>
<td>129</td>
<td>1</td>
</tr>
<tr>
<td>97</td>
<td>1</td>
</tr>
<tr>
<td>84</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>39</td>
<td>0</td>
</tr>
<tr>
<td>60</td>
<td>0</td>
</tr>
<tr>
<td>65</td>
<td>1</td>
</tr>
<tr>
<td>94</td>
<td>1</td>
</tr>
<tr>
<td>86</td>
<td>1</td>
</tr>
<tr>
<td>64</td>
<td>1</td>
</tr>
</tbody>
</table>


Is there a statistically significant difference in mean palatability score between the two texture levels? Show all four steps in a significance test when answering this question.

E61. Is there sufficient evidence from the random samples of heart rates for men and women under normal conditions in Display 9.66 to say that the mean heart rates differ for the two groups? Analyze the data in two different ways (confidence interval and test of significance) before coming to your conclusion, and then state your conclusion carefully.

### Display 9.66 Heart rates.

<table>
<thead>
<tr>
<th>Heart Rate (beats per minute)</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>74</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>69</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>76</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>78</td>
<td>69</td>
<td></td>
</tr>
<tr>
<td>78</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td>86</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>84</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>71</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>74</td>
<td></td>
</tr>
</tbody>
</table>


E62. The data on mercury content of fish in the lakes of Maine in E48 on page 613 are augmented in Display 9.67 by two other variables: whether the lake is formed behind a dam and the oxygen content of the water.
Mercury content of fish in Maine lakes.

<table>
<thead>
<tr>
<th>Mercury (ppm)</th>
<th>Lake Type</th>
<th>Dam (1 = yes; 0 = no)</th>
<th>Mercury (ppm)</th>
<th>Lake Type</th>
<th>Dam (1 = yes; 0 = no)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.050</td>
<td>2</td>
<td>1</td>
<td>0.450</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>0.230</td>
<td>2</td>
<td>1</td>
<td>1.120</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>0.100</td>
<td>3</td>
<td>0</td>
<td>0.320</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>0.770</td>
<td>2</td>
<td>1</td>
<td>0.370</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>0.910</td>
<td>2</td>
<td>1</td>
<td>0.540</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>0.250</td>
<td>2</td>
<td>1</td>
<td>0.860</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>0.130</td>
<td>1</td>
<td>1</td>
<td>0.770</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>0.290</td>
<td>2</td>
<td>0</td>
<td>0.670</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>0.410</td>
<td>3</td>
<td>1</td>
<td>0.600</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>0.210</td>
<td>3</td>
<td>0</td>
<td>0.680</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>0.940</td>
<td>2</td>
<td>0</td>
<td>0.220</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>0.360</td>
<td>1</td>
<td>1</td>
<td>0.470</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>1.220</td>
<td>2</td>
<td>0</td>
<td>0.370</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>0.240</td>
<td>1</td>
<td>1</td>
<td>0.290</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>0.900</td>
<td>3</td>
<td>0</td>
<td>0.430</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2.500</td>
<td>2</td>
<td>1</td>
<td>0.160</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0.340</td>
<td>3</td>
<td>0</td>
<td>0.490</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>


a. Some environmentalists claim that dams are a cause of high mercury levels. Is there sufficient evidence to conclude that the mean mercury level of the lakes behind dams is greater than the mean mercury level of the lakes without dams?

b. Type 1 lakes are *oligotrophic* (balanced between decaying vegetation and living organisms), type 2 lakes are *eutrophic* (high decay rate and little oxygen), and type 3 lakes are *mesotrophic* (between the other two states). Estimate the difference in mean mercury level for type 1 and type 2 lakes?

c. Is there sufficient evidence to conclude that the mean mercury level differs for lakes of type 2 compared to lakes of type 3?

E63. Refer to Kelly’s hamster experiment in E53 on page 632.

a. What kind of error could Kelly be making in a significance test of the difference in mean enzyme concentration?

b. What should Kelly do in a similar experiment if she wants to reduce the probability of an error of this type?

c. What about this situation gave Kelly power to reject the null hypothesis even with such small treatment group sizes?

E64. Refer to the taste test experiment in E60.

a. What kind of error might investigators be making in this significance test?

b. What should the investigators do in a similar experiment if they want to reduce the probability of this type of error?

E65. As you read in E1 and P27, some students recruited 30 volunteers to attempt to walk the length of a football field while blindfolded. Each volunteer began at the middle of one goal line and was asked to walk to the opposite goal line, a distance of 100 yards. Display 9.68 shows the distance at which the volunteers crossed a sideline of the field and whether they were male or female.

Display 9.68  Boxplot and stemplot of yard-line data, classified by gender.

<table>
<thead>
<tr>
<th>Yard Line</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>6 2</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>8 1 0 0</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>5 4 3 0 0</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>

Display 9.68  Boxplot and stemplot of yard-line data, classified by gender.

a. Is this study more like an experiment or more like a survey?
b. Conduct a test of significance for the difference in the mean number of yards walked for males and females. Use the summary statistics in the printout in Display 9.69.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Gender</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>TrMean</th>
<th>StDev</th>
<th>SEMean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yard Lin</td>
<td>F</td>
<td>15</td>
<td>51.40</td>
<td>50.00</td>
<td>50.38</td>
<td>13.40</td>
<td>3.46</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>15</td>
<td>65.33</td>
<td>68.00</td>
<td>65.23</td>
<td>14.10</td>
<td>3.64</td>
</tr>
</tbody>
</table>

Display 9.69  Summary statistics of yard-line data for males and females.

c. Display 9.70 shows the relationship between the height of the volunteer, in inches, and the number of yards he or she walked before crossing a sideline. What lurking variable can help explain your result in part b?

Display 9.70  Scatterplot of yard-line data, categorized by gender.

E66. As you saw in E43 on page 611, a study to compare two insurance companies on length of stay (LOS) for pediatric asthma patients randomly sampled 393 cases from Insurer A. Summary statistics and a histogram for the data are shown in Display 9.71.

<table>
<thead>
<tr>
<th>Summary of Insurer A: Length of Stay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>SD</td>
</tr>
</tbody>
</table>


An independent random sample of 396 cases from Insurer B gave the results on length of stay summarized in Display 9.72.

<table>
<thead>
<tr>
<th>Summary of Insurer B: Length of Stay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>SD</td>
</tr>
</tbody>
</table>

Display 9.72  Summary and data plot for lengths of hospital stays for Insurer B.

a. Estimate the difference between the mean lengths of stay for the two insurance companies in a 95% confidence interval. Is there evidence of a difference between the population means for the two companies?

b. Many other variables could contribute to the difference in mean length of stay for the two insurers. One is the number of full-time staff per bed. The sample means for this variable are 4.63 for Insurer A and 6.13 for Insurer B. The respective sample standard deviations are 1.70 and 2.40. Is this difference in means statistically significant? If so, provide a practical explanation of how this difference might be related to the difference in mean length of stay.

c. Another contributing variable is the percentage of private hospitals versus public hospitals among the patients in each insurer’s sample. For Insurer A, 93.6% of the sampled patients are in...
private hospitals; for Insurer B, that percentage is 73.0%. Is this a statistically significant difference? If so, how might this difference be related to the difference in mean length of stay?

E67. There is much controversy about whether coaching programs improve scores on SAT exams by more than a minimal amount. Although this exercise won’t settle the argument, it does present a practical use of a test of significance.

The College Board reports that coaching programs improve SAT mathematics scores by about 25 points on average. A sample of 50 students in a “control” group who took the math portion of the SAT as juniors and then again as seniors had an average gain of 13 points (reported to be about the national average gain without coaching), with standard deviation of about 30 points. A sample of 9 students who were coached between their two exams had an average gain of 60 points, with a standard deviation of 42. [Source: Jack Kaplan, “An SAT Coaching Program That Works,” Chance 15, no. 1 (2002): 12–17.]

a. Is there evidence that the difference in mean point gain between the coached group and the control group exceeds the 25 points that is to be expected (according to the College Board)?

b. Another sample of 12 students from a coaching program had an average gain of 73 points, with a standard deviation of 42. Is there evidence that the difference in mean point gain here exceeds the 25 points that is to be expected?

c. What further questions would you like to ask about this study?

E68. An important factor in the performance of a pharmaceutical product is how fast the product dissolves in vivo (in the body). This is measured by a dissolution test, which yields the percentage of the label strength (%LS) released after certain elapsed times. Laboratory tests of this type are conducted in vessels that simulate the action of the stomach. Display 9.73 shows %LS at certain time intervals for an analgesic (painkiller) tested in laboratories in New Jersey and Puerto Rico. Time is measured in minutes.


<table>
<thead>
<tr>
<th>New Jersey Elapsed Time (min)</th>
<th>Vessel No. (%LS)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>V1</td>
<td>V2</td>
<td>V3</td>
<td>V4</td>
<td>V5</td>
<td>V6</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>5</td>
<td>10</td>
<td>2</td>
<td>7</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>40</td>
<td>72</td>
<td>79</td>
<td>81</td>
<td>70</td>
<td>72</td>
<td>73</td>
</tr>
<tr>
<td>60</td>
<td>96</td>
<td>99</td>
<td>93</td>
<td>95</td>
<td>96</td>
<td>99</td>
</tr>
<tr>
<td>120</td>
<td>99</td>
<td>99</td>
<td>96</td>
<td>100</td>
<td>98</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Puerto Rico Elapsed Time (min)</th>
<th>Vessel No. (%LS)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>V1</td>
<td>V2</td>
<td>V3</td>
<td>V4</td>
<td>V5</td>
<td>V6</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>12</td>
<td>7</td>
<td>3</td>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td>40</td>
<td>65</td>
<td>66</td>
<td>71</td>
<td>70</td>
<td>74</td>
<td>69</td>
</tr>
<tr>
<td>60</td>
<td>95</td>
<td>99</td>
<td>98</td>
<td>94</td>
<td>90</td>
<td>92</td>
</tr>
<tr>
<td>120</td>
<td>100</td>
<td>102</td>
<td>98</td>
<td>99</td>
<td>97</td>
<td>100</td>
</tr>
</tbody>
</table>

- a. Use a 90% confidence interval to estimate the difference in mean %LS at 40 minutes for New Jersey compared to Puerto Rico.
- b. Use a 90% confidence interval to estimate the difference in mean %LS at 60 minutes for New Jersey compared to Puerto Rico.
- c. Good manufacturing practices call for “equivalence limits” of 15 percentage points for dissolution percentages below 90% and 7 percentage points for dissolution percentages above 90%. That is, if the 90% confidence interval for the mean difference is within the equivalence limits (within an interval from $-15\%$ to $+15\%$ for the lower percentages), then the two sets of results are accepted as equivalent. Will the results in parts a and b be accepted as equivalent at 40 minutes? At 60 minutes?
Paired Comparisons

Now that you’ve seen the methods for comparing means, it’s time to put them to work. In this section, you’ll see confidence intervals and significance tests in action. Keep in mind that a t-test is no smarter than a chainsaw. Neither has any brains of its own. A chainsaw can’t tell whether it’s cutting an old dead tree into firewood or turning a valuable antique table into scrapwood. A t-test is every bit as oblivious. The difference between thoughtful and careless use is up to you, the operator. This final section looks at four issues on which you need to be clear in order to use your statistical tools with care.

- Do you really have two independent samples, or do you have only one sample of paired data?
- What if shapes aren’t normal?
- Is it meaningful to compare means?
- Does your inference have the chance it needs?

Two Independent Samples, or Paired Observations?

One of the recurring themes in statistics is how important it is to pay attention to data production. This theme was developed in Chapter 4, where the entire emphasis was on designing studies. Now the theme returns in the context of inference for means, where one of the key questions is “Do I have paired data or independent samples?” Activity 9.5a will help you see how pairing changes the standard error.

Hand Spans

What you’ll need: a ruler marked in millimeters

Detective Sherlock Holmes amazed a man by relating “obvious facts” about him, such as that he had at some time done manual labor: “How did you know, for example, that I did manual labour? It is as true as gospel, for I began as a ship's carpenter.” Sherlock replied, “Your hands, my dear sir. Your right hand is quite a size larger than your left. You have worked with it, and the muscles are more developed.” [Source: Sir Arthur Conan Doyle, *The Adventures of Sherlock Holmes*, ed. Richard Lancelyn Green, Oxford World Classics (Oxford and New York, 1988).] In fact, people’s right hands tend to be bigger than their left, even if they are left-handed and even if they haven’t done manual labor, but the difference is small. To detect it, you will have to design your study carefully.

(continued)
1. Measure your left and right hand spans, in millimeters. (An easy way to do this is to spread your hand as wide as possible, place it directly on a ruler, and get the distance between the end of your little finger and the end of your thumb.) Record the data for each student in your class in a table with a column for the left hand span and another column for the right. There should be one row for each person.

2. For each row in the table, calculate the difference, right — left. Find the mean of the differences and the standard error of the mean difference using the formula \( s_d = s/\sqrt{n} \).

3. Now make a new table, but this time randomize the order of the right hand spans so that people’s left hand spans are no longer matched with their right. Then repeat step 2 with the new table.

4. Finally, treat the left hand spans and right hand spans as independent samples. Calculate the difference between the two sample means and the standard error of that difference using the formula

\[
\frac{s_{x_R - x_L}}{s_{x_R} \sqrt{\frac{1}{n_R}} + s_{x_L} \sqrt{\frac{1}{n_L}}}
\]

5. Compare the standard errors from steps 2, 3, and 4. Which is smallest? Which two are most nearly the same size?

6. Suppose you make scatterplots of the data in your two tables from steps 1 and 3, with the data for the left hand spans on the horizontal axis. Which would you expect to have the higher correlation? Why? Make the scatterplots, and calculate the correlations to check your answer.

The lesson of Activity 9.5a is this: Paired observations can greatly reduce variation over independent samples and produce a much more powerful test and a more precise confidence interval estimate of the true mean difference. The reduction in variation is greatest when the underlying measurements vary greatly from pair to pair but the differences do not.

In Activity 4.4a, you compared sitting and standing pulse rates using three different designs. In the completely randomized design, you randomly selected half of the class to sit and the other half to stand. These data can be analyzed using the techniques for two independent samples, as in Section 9.4. In the matched pairs design, you matched pairs of students on a preliminary measure of pulse rate. Then you randomly assigned sitting to one student in each pair and standing to the other. In the repeated measures design, you had each person sit and stand, with the order randomly assigned. The data from the last two designs should not be analyzed using the techniques for two independent samples. In this section, you will learn how to analyze these data using the differences in pulse rate for each pair. Display 9.74 (on the next page) shows the data for pulse rates, in beats per minute, from a class that worked through this activity.
### Completely Randomized Design

<table>
<thead>
<tr>
<th>Sit</th>
<th>Stand</th>
</tr>
</thead>
<tbody>
<tr>
<td>70</td>
<td>66</td>
</tr>
<tr>
<td>88</td>
<td>82</td>
</tr>
<tr>
<td>82</td>
<td>86</td>
</tr>
<tr>
<td>88</td>
<td>102</td>
</tr>
<tr>
<td>66</td>
<td>62</td>
</tr>
<tr>
<td>70</td>
<td>70</td>
</tr>
<tr>
<td>72</td>
<td>50</td>
</tr>
<tr>
<td>86</td>
<td>62</td>
</tr>
<tr>
<td>74</td>
<td>56</td>
</tr>
<tr>
<td>86</td>
<td>104</td>
</tr>
<tr>
<td>80</td>
<td>86</td>
</tr>
<tr>
<td>46</td>
<td>86</td>
</tr>
<tr>
<td>54</td>
<td>80</td>
</tr>
<tr>
<td>86</td>
<td>96</td>
</tr>
</tbody>
</table>

### Matched Pairs Design

<table>
<thead>
<tr>
<th>Sit</th>
<th>Stand</th>
<th>Difference (stand – sit)</th>
</tr>
</thead>
<tbody>
<tr>
<td>62</td>
<td>68</td>
<td>6</td>
</tr>
<tr>
<td>74</td>
<td>78</td>
<td>4</td>
</tr>
<tr>
<td>82</td>
<td>80</td>
<td>–2</td>
</tr>
<tr>
<td>88</td>
<td>92</td>
<td>4</td>
</tr>
<tr>
<td>82</td>
<td>58</td>
<td>–24</td>
</tr>
<tr>
<td>66</td>
<td>96</td>
<td>30</td>
</tr>
<tr>
<td>64</td>
<td>72</td>
<td>8</td>
</tr>
<tr>
<td>84</td>
<td>100</td>
<td>16</td>
</tr>
<tr>
<td>72</td>
<td>82</td>
<td>10</td>
</tr>
<tr>
<td>82</td>
<td>76</td>
<td>–6</td>
</tr>
<tr>
<td>80</td>
<td>92</td>
<td>12</td>
</tr>
<tr>
<td>72</td>
<td>74</td>
<td>2</td>
</tr>
<tr>
<td>64</td>
<td>60</td>
<td>–4</td>
</tr>
<tr>
<td>62</td>
<td>58</td>
<td>–4</td>
</tr>
</tbody>
</table>

### Matched Pairs Design

<table>
<thead>
<tr>
<th>Sit</th>
<th>Stand</th>
<th>Difference (stand – sit)</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>64</td>
<td>4</td>
</tr>
<tr>
<td>70</td>
<td>72</td>
<td>2</td>
</tr>
<tr>
<td>72</td>
<td>76</td>
<td>4</td>
</tr>
<tr>
<td>78</td>
<td>82</td>
<td>4</td>
</tr>
<tr>
<td>80</td>
<td>92</td>
<td>12</td>
</tr>
<tr>
<td>84</td>
<td>98</td>
<td>14</td>
</tr>
<tr>
<td>60</td>
<td>68</td>
<td>8</td>
</tr>
<tr>
<td>62</td>
<td>64</td>
<td>2</td>
</tr>
<tr>
<td>66</td>
<td>70</td>
<td>4</td>
</tr>
<tr>
<td>72</td>
<td>86</td>
<td>14</td>
</tr>
<tr>
<td>82</td>
<td>100</td>
<td>18</td>
</tr>
<tr>
<td>74</td>
<td>80</td>
<td>6</td>
</tr>
<tr>
<td>50</td>
<td>58</td>
<td>8</td>
</tr>
<tr>
<td>52</td>
<td>54</td>
<td>2</td>
</tr>
<tr>
<td>64</td>
<td>66</td>
<td>2</td>
</tr>
<tr>
<td>70</td>
<td>76</td>
<td>6</td>
</tr>
<tr>
<td>76</td>
<td>86</td>
<td>10</td>
</tr>
<tr>
<td>70</td>
<td>88</td>
<td>18</td>
</tr>
<tr>
<td>88</td>
<td>96</td>
<td>8</td>
</tr>
<tr>
<td>80</td>
<td>86</td>
<td>6</td>
</tr>
<tr>
<td>54</td>
<td>56</td>
<td>2</td>
</tr>
<tr>
<td>68</td>
<td>82</td>
<td>14</td>
</tr>
<tr>
<td>86</td>
<td>96</td>
<td>10</td>
</tr>
<tr>
<td>68</td>
<td>74</td>
<td>6</td>
</tr>
<tr>
<td>68</td>
<td>80</td>
<td>12</td>
</tr>
<tr>
<td>48</td>
<td>58</td>
<td>10</td>
</tr>
<tr>
<td>64</td>
<td>72</td>
<td>8</td>
</tr>
<tr>
<td>74</td>
<td>94</td>
<td>20</td>
</tr>
</tbody>
</table>

### Summary Statistics

<table>
<thead>
<tr>
<th>CRD</th>
<th>MPD</th>
<th>RMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sit</td>
<td>Stand</td>
<td>Sit</td>
</tr>
<tr>
<td>74.86</td>
<td>77.71</td>
<td>73.86</td>
</tr>
<tr>
<td>13.00</td>
<td>17.04</td>
<td>9.13</td>
</tr>
</tbody>
</table>

**Display 9.74** Pulse measurements, in beats per minute, from a class experiment.
Completely Randomized Design (Two Independent Samples)

You will now work through the analysis of each of the three sets of experimental results, beginning with the completely randomized design. Display 9.75 shows the relationship between the sitting and standing pulse rates for the arbitrary pairing of values from the two independent samples in the completely randomized design table in Display 9.74. The correlation is near 0, as it should be, because these are independent measurements taken in arbitrary order.

To analyze the data from this completely randomized design, use the methods of Section 9.4. A 95% confidence interval \((df = 24.31)\) for the difference between the mean pulse rates for sitting and standing is given by

\[
(\bar{x}_1 - \bar{x}_2) \pm t^* \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (77.71 - 74.86) \pm t^* \cdot \sqrt{\frac{17.04^2}{14} + \frac{13.00^2}{14}}
\]

or \(-8.96 < \mu_{\text{stand}} - \mu_{\text{sit}} < 14.67\). This interval overlaps 0, so you can’t conclude that one of these treatments would produce a higher mean than the other if every subject were given both treatments.

Matched Pairs Design

Display 9.76 shows the relationship between the sitting and standing pulse rates for the matched pairs design.

\[
(\bar{x}_1 - \bar{x}_2) \pm t^* \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (77.71 - 74.86) \pm t^* \cdot \sqrt{\frac{17.04^2}{14} + \frac{13.00^2}{14}}
\]
This plot gives a hint of a linear trend \((r = 0.48)\), as it should, because these are dependent measurements based on pairing people with similar resting pulse rates.

The matched pairs design has dependent observations within a pair, so the two-sample \(t\)-procedure is not a valid option. You can, however, look at the differences between the standing and sitting pulse rates for each pair and estimate the mean difference with a one-sample procedure. Observe in Display 9.74 that the difference between the sample means, \(77.57 - 73.86\), is the same as the mean of the differences, \(3.71\), so the latter is a legitimate estimator of the true difference between the population means. Thus, the two sets of measurements are reduced to one set of differences. The summary statistics for the observed differences are

\[
\text{Mean: } \bar{d} = 3.71 \quad \text{Standard deviation: } s_d = 12.38
\]

With 13 degrees of freedom, the critical value, \(t^*\), for a 95% confidence interval is 2.160. The confidence interval estimate of the true mean difference is

\[
\bar{d} \pm t^* \cdot \frac{s_d}{\sqrt{n}} \quad \text{or} \quad 3.71 \pm 2.160 \cdot \frac{12.38}{\sqrt{14}}
\]

which yields the interval \((-3.44, 10.86)\). Any value of the true mean difference between standing and sitting pulse rates in this interval could have produced the observed mean difference as a reasonably likely outcome. This interval includes 0, so there is not sufficient evidence to say that the standing mean would differ from the sitting mean if all subjects were measured under both conditions. Although this interval overlaps 0, 0 is proportionally closer to the endpoint than it is in the confidence interval for the completely randomized design.

Repeating Measures Design

The scatterplot for the data from the repeated measures design (Display 9.77) shows a strong linear trend (correlation 0.93). These are paired measurements from the same person and should be highly correlated.

\[
\text{Display 9.77}\quad \text{Scatterplot of sitting versus standing pulse rates, in beats per minute, in a repeated measures design.}
\]

A 95% confidence interval for the mean difference \((df = 27)\) is

\[
\bar{d} \pm t^* \cdot \frac{s_d}{\sqrt{n}} \quad \text{or} \quad 8.36 \pm 2.052 \cdot \frac{5.28}{\sqrt{28}}
\]
or (6.31, 10.41). This estimate of the mean does not overlap 0, so the evidence supports the conclusion that the mean standing pulse rate is higher than the mean sitting pulse rate.

**Checking Normality**

You might be wondering why we did not check conditions by looking at a plot of the data. We will do that now. For the analyses of differences, it is the differences, not the two original samples, that must come from an approximately normal distribution. Display 9.78 provides boxplots of these differences for both the matched pairs design and the repeated measures design. They show that there is no obvious reason to be concerned about non-normality in either case.

![Boxplots of differences in pulse rates.](image)

**A Confidence Interval for Paired Observations**

The box summarizes how to construct a confidence interval for the mean difference.

---

**A Confidence Interval for the Mean Difference from Paired Observations**

**Check conditions.**

You must check three conditions:

- *Randomness.* In the case of a survey, you must have a random sample from one population, where a “unit” might consist of, say, a pair of twins or the same person’s two feet. In the case of an experiment, the treatments must have been randomly assigned within each pair. If the same subject is assigned both treatments, the treatments must have been assigned in random order.

- *Normality.* The differences must look like it’s reasonable to assume that they came from a normally distributed population or the sample size must be large enough that the sampling distribution of the sample mean difference is approximately normal. The 15/40 guideline can be applied to the differences.

- *Sample Size.* In the case of a sample survey, the size of the population of differences should be at least ten times as large as the sample size.

**Do computations.**

A confidence interval for the mean difference, \( \mu_d \), is given by

\[
\bar{d} \pm t^* \cdot \frac{s_d}{\sqrt{n}}
\]

(continued)
where \( n \) is the sample size, \( \bar{d} \) is the mean of the differences between pairs of measurements, \( s_d \) is the sample standard deviation, and \( t^* \) depends on the confidence level desired and the degrees of freedom, \( df = n - 1 \).

**Give interpretation in context.**
An interpretation is of this form: “I am 95% confident that the mean of the population of differences, \( \mu_d \), is in this confidence interval.”

Of course, when you interpret a confidence interval, you will do it in context, describing the population you are talking about.

---

**A Significance Test for Paired Observations**

You could also decide whether standing increases the mean pulse rate by doing a test of significance. The next example illustrates the components of such a test.

**Example: Testing the Mean Difference for the Repeated Measures Design**

Use the data in the repeated measures design of Display 9.74 to test the research hypothesis that standing increases the mean pulse rate.

**Solution**

Because the order in which each subject receives the two treatments is randomized, you can treat this as a random assignment of treatments to subjects. Although Display 9.78 shows that the distribution of differences is not quite symmetric, there is no reason to rule out the normal distribution as a possible model for producing these differences.

The hypotheses are

\[
H_0: \mu_d = 0 \quad \text{versus} \quad H_a: \mu_d > 0
\]

Here \( \mu_d \) is the theoretical mean difference between standing and sitting pulse rates for this group of subjects if each subject could have received each treatment in both orders.

The test statistic is

\[
t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} = \frac{8.36 - 0}{5.28/\sqrt{28}} = 8.38
\]

With 27 degrees of freedom, the \( P \)-value for this large \( t \)-statistic is essentially 0.

The very small \( P \)-value indicates that there is sufficient evidence to conclude that the mean difference in pulse rates is positive. This implies that the mean pulse rate is higher for persons standing than it is for those same persons sitting. The result applies only to the people (subjects) in this experiment and cannot be generalized to other people based on these data alone.
The procedure is summarized in this box.

Components of a Test of the Mean Difference Based on Paired Observations

1. **Name the test and check conditions.** For a test of significance of a mean difference based on paired observations, the method requires the same conditions as those for a confidence interval. You must check three conditions:
   - **Randomness.** In the case of a survey, you must have a random sample from one population, where a “unit” might consist of, say, a pair of twins or the same person’s two feet. In the case of an experiment, the treatments must have been randomly assigned within each pair. If the same subject is given both treatments, the treatments must have been assigned in random order.
   - **Normality.** The differences must look like it’s reasonable to assume that they came from a normally distributed population or the sample size must be large enough that the sampling distribution of the sample mean difference is approximately normal. The 15/40 guideline can be applied to the differences.
   - **Sample Size.** In the case of sample surveys, the size of the population of differences should be at least ten times as large as the number of differences in the sample.

2. **State the hypotheses.** The null hypothesis is, ordinarily, that in the entire population the mean of the differences is 0. In symbols, \( H_0: \mu_d = 0 \).
   
   There are three forms of the alternative or research hypothesis:
   - \( H_a: \mu_d \neq 0 \)
   - \( H_a: \mu_d < 0 \)
   - \( H_a: \mu_d > 0 \)

3. **Compute the test statistic, find the \( P \)-value, and draw a sketch.** Compute the difference between the mean difference, \( \bar{d} \), from the sample and the hypothesized difference \( \mu_d \), and then divide by the estimated standard error:

   \[
   t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}}
   \]

   The \( P \)-value is based on \( n - 1 \) degrees of freedom, where \( n \) is the number of differences.

4. **Write your conclusion linked to your computations and in the context of the problem.** If you are using fixed-level testing, reject the null hypothesis if your \( P \)-value is less than the level of significance, \( \alpha \). If the \( P \)-value is greater than or equal to \( \alpha \), do not reject the null hypothesis. (If you are not given a value of \( \alpha \), you can assume that \( \alpha \) is \( 0.05 \).) Write a conclusion that relates to the situation and includes an interpretation of your \( P \)-value.
Two Independent Samples, or Paired Observations?

D16. Based on what you recall from Chapter 4 and what you saw in the analysis of the pulse rate data in this section, discuss the relative merits of completely randomized, matched pairs, and repeated measures designs. Under what conditions will the analysis of differences between pairs pay bigger dividends than the analysis of differences of independent means?

What If the Differences Don’t Appear Normal?

You discovered in Section 9.3 that the true capture rate of a nominal 95% confidence interval is less than 95% if the population is skewed and the sample size is not large enough for the sampling distribution of \( \bar{x} \) to be approximately normal. The same is true of the confidence interval for the mean difference. If the sample differences appear to come from a skewed distribution, it is usually better to transform the original data to a new scale that reduces the skewness.

Example: Transforming Paired Data

The data in Display 9.79 are counts of leprosy bacilli colonies at specified sites on human subjects. The columns labeled \( B \) and \( A \) show the counts on the same subjects before and after an antiseptic was applied. The order of the measurements in each pair cannot be randomized, but these subjects were randomly selected from a larger group. The research question is “Is the antiseptic effective in reducing bacteria counts?”

<table>
<thead>
<tr>
<th>( B )</th>
<th>( A )</th>
<th>( B - A )</th>
<th>( \sqrt{B} - \sqrt{A} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0</td>
<td>6</td>
<td>2.45</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>4</td>
<td>1.04</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>4</td>
<td>0.91</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>7</td>
<td>1.83</td>
</tr>
<tr>
<td>18</td>
<td>18</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>4</td>
<td>0.83</td>
</tr>
<tr>
<td>19</td>
<td>14</td>
<td>5</td>
<td>0.62</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>-1</td>
<td>-0.17</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>4</td>
<td>1.24</td>
</tr>
<tr>
<td>15</td>
<td>9</td>
<td>6</td>
<td>0.87</td>
</tr>
</tbody>
</table>


Solution

Check conditions.

This particular group of subjects is a random sample taken from a larger group, so the randomness condition is satisfied. But the distributions of the original counts and the distribution of their differences don’t appear normal, as you can see in Display 9.80. Counts of this type are notorious for having skewed
distributions because they must be non-negative but can get very large. (Think of counting the number of insects on each plant in a garden.) Display 9.80 also shows that these particular counts have positively skewed distributions, whereas their differences are skewed in the negative direction. A transformation is needed. Although the condition is that the distribution of differences must be normal, statisticians typically transform the original values rather than the differences.

Transform the original values, not the differences.

\[
\begin{align*}
B - A & \quad * \quad * \\
A & \\
B & \\
\end{align*}
\]

Display 9.80  Boxplots of the original counts, B and A, and of the differences, B - A.

If you take the square root of each of the counts in B and A, you get the more symmetric boxplots in Display 9.81. The square root transformation typically does a good job of making the distributions of counts more symmetric.

\[
\begin{align*}
\sqrt{B} - \sqrt{A} & \quad * \\
\sqrt{A} & \\
\sqrt{B} & \\
\end{align*}
\]

Display 9.81  Boxplots of the square roots of the counts and of the differences of the square roots of the counts.

The purpose of applying the antiseptic was, of course, to reduce the bacteria counts. Evidence that the treatment was effective can be produced through a \( t \)-test of the differences of the square roots. The data in Display 9.79 have the “after” counts subtracted from the “before” counts, so the hypotheses are

\[
H_0: \mu_d = 0 \quad \text{and} \quad H_a: \mu_d > 0
\]

where \( \mu_d \) is the mean of the population of differences of the square roots of the number of bacteria colonies before and after the antiseptic is applied.

The statistical summary for the differences of the square roots is

Mean of differences of square roots: \( \bar{d} = 0.96 \)

Standard deviation of differences of square roots: \( s_d = 0.77 \)
The formula for the test statistic has the same pattern as in the $t$-test for a single mean:

\[
t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}}
\]

\[
= \frac{0.96 - 0}{0.77/\sqrt{10}}
\]

\[
= 3.943
\]

With $df = 9$, the $P$-value is 0.0017.

The small $P$-value, 0.0017, gives ample evidence to support the conclusion that the bacteria counts were lower after the antiseptic was applied; that is, the true mean difference is positive. However, because there was no randomization in the order of treatments, perhaps you can’t attribute the decrease to the antiseptic. There is always the possibility that the bacteria count might have gone down over time anyway. (See E78 on page 659 to find out how the experimenters eliminated this possibility.)

**DISCUSSION**

**What If the Differences Don’t Appear Normal?**

D17. In analyzing differences taken on skewed data, the data could be transformed before the differences are calculated, or the differences could be calculated first and then transformed. Does the order of doing these tasks matter? Why or why not?

**Is It Meaningful to Compare Means?**

“When your only tool is a hammer, everything looks like a nail.” By now you have quite a variety of statistical tools, but it’s still worth reminding yourself to be thoughtful about when and how you use them. Before you start any computations, always ask yourself, “What is the question of interest?” Then ask whether it makes sense to try to answer that question by comparing means.

**Example: Comparing Aptitude Tests**

Business firms often give aptitude tests to job applicants to see if they have the necessary skills and interests to adapt well to any training they might have to complete. Suppose a company desiring to use such a test has two versions it wants to compare, Test A and Test B. To see how well the exams predict aptitude for a job, the company decides to try both exams on a sample of recently hired employees who have completed the training (and whose aptitude they now know). Is the company interested in comparing the mean scores of the two groups?

**Solution**

Probably not. The two tests might be graded on different scales, so a comparison of means would be meaningless anyway. More important, the company really is interested in how well the tests select the applicants with the greatest aptitude, so
they might want to compare, say, the top 10% of test takers on each exam to see if they really are the employees with the greatest aptitude. In general, the variability of the test scores is more important than the mean. Those with strong aptitude should score high; those with weak aptitude should score low, with enough variability in scores so that the two groups are substantively different.

Does Your Inference Have the Chance It Needs?

In statistics, exploratory methods look for patterns but make no assumptions about the process that created the data; with exploratory methods, what you see is what you get. Inference can deliver more than exploration, going beyond simply saying “Here are some interesting patterns.” For inference to be justified, however, you need the right kind of data. Provided your numbers come from random samples or randomized experiments, you can use probability theory and the predictable regularities of chancelike behavior to draw conclusions not only about the data you see but also about the unseen population from which they came or about the treatments that were assigned to the experimental units.

Is It Meaningful to Compare Means??

D18. Make a list of ways your data can fall short of the requirements for statistical inference on means.

Summary 9.5: Paired Comparisons

Textbook illustrations of statistical procedures are always neater than the real-world applications of those same procedures. When confronted with a real problem involving two samples that has not been “sanitized” for textbook use, you should ask some key questions.

- “Are these paired observations, or two independent samples?” If the data are paired, they are likely to be correlated, and the two-sample procedures are not the correct method. The better method is to analyze the differences between pairs as a single random sample using the methods of this section.
- “How skewed is the population from which these differences came?” If you have paired differences, check to see if it’s reasonable to assume that the differences are a random sample taken from a normally distributed population. If not, try a transformation on the original values (not on the differences).
- “Is this a meaningful comparison?” Anyone armed with statistical software can compare the means of any two sets of data. But some might result in comparing the proverbial apples and oranges, and some might involve comparing samples that were constructed to have different means in the first place.
• “Is there any chance mechanism underlying the selection of these data?”
Statistical inference is based on probability theory, and the procedures
work only under the condition that the data come from random samples or
randomized experiments. You can then draw conclusions not only about the
data you see but also about the unseen population from which they came or
about the reason why the treatment groups differ.

Practice
Two Independent Samples, or Paired
Observations?
P33. Display 9.82 shows a set of data on pulse
rates, in beats per minute, from another
class experiment employing the same three
designs as in Display 9.74 on page 642.

Display 9.83 shows the sitting pulse rate
plotted against the standing pulse rate for
each of the three designs. For which design
is the relationship strongest? Weakest?
Explain why that should be the case.

<table>
<thead>
<tr>
<th>Completely Randomized Design</th>
<th>Matched Pairs Design</th>
<th>Repeated Measures Design</th>
<th>Summary Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sit</td>
<td>Stand</td>
<td>Sit</td>
<td>Stand</td>
</tr>
<tr>
<td>78</td>
<td>76</td>
<td>74</td>
<td>78</td>
</tr>
<tr>
<td>64</td>
<td>68</td>
<td>76</td>
<td>74</td>
</tr>
<tr>
<td>50</td>
<td>82</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>58</td>
<td>80</td>
<td>96</td>
<td>90</td>
</tr>
<tr>
<td>50</td>
<td>68</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td>70</td>
<td>64</td>
<td>74</td>
<td>74</td>
</tr>
<tr>
<td>70</td>
<td>58</td>
<td>62</td>
<td>70</td>
</tr>
<tr>
<td>64</td>
<td>90</td>
<td>68</td>
<td>66</td>
</tr>
<tr>
<td>66</td>
<td>72</td>
<td>64</td>
<td>74</td>
</tr>
<tr>
<td>72</td>
<td>78</td>
<td>60</td>
<td>80</td>
</tr>
<tr>
<td>80</td>
<td>60</td>
<td>56</td>
<td>58</td>
</tr>
<tr>
<td>56</td>
<td>100</td>
<td>52</td>
<td>52</td>
</tr>
<tr>
<td>58</td>
<td>80</td>
<td>80</td>
<td>88</td>
</tr>
<tr>
<td>68</td>
<td>84</td>
<td>70</td>
<td>66</td>
</tr>
</tbody>
</table>

Display 9.82 Data tables and boxplots of pulse rate data from another class experiment.
Display 9.83 Scatterplots of pulse rate data.

b. Perform a significance test that standing increases heart rate for the completely randomized design.

c. Perform a significance test that standing increases heart rate for the matched pairs design.

d. Perform a significance test that standing increases heart rate for the repeated measures design.

e. Compare the $P$-values for the three designs in parts b–d. Do they show the pattern you would expect?

What If the Differences Don’t Appear Normal?

P34. Fish or fowl, which are smarter? Display 9.84 lists brain weights of a sample of birds and fish selected from the data in Display 9.31 on page 603.

Display 9.84 Brain weights of a sample of birds and fish.

<table>
<thead>
<tr>
<th>Birds</th>
<th>Brain Weight (g)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canary</td>
<td>0.85</td>
</tr>
<tr>
<td>Crow</td>
<td>9.30</td>
</tr>
<tr>
<td>Flamingo</td>
<td>8.05</td>
</tr>
<tr>
<td>Loon</td>
<td>6.12</td>
</tr>
<tr>
<td>Pheasant</td>
<td>3.29</td>
</tr>
<tr>
<td>Pigeon</td>
<td>2.69</td>
</tr>
<tr>
<td>Vulture</td>
<td>19.60</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fish</th>
<th>Brain Weight (g)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barracuda</td>
<td>3.83</td>
</tr>
<tr>
<td>Brown trout</td>
<td>0.57</td>
</tr>
<tr>
<td>Catfish</td>
<td>1.84</td>
</tr>
<tr>
<td>Mackerel</td>
<td>0.64</td>
</tr>
<tr>
<td>Northern trout</td>
<td>1.23</td>
</tr>
<tr>
<td>Salmon</td>
<td>1.26</td>
</tr>
<tr>
<td>Tuna</td>
<td>3.09</td>
</tr>
</tbody>
</table>

a. When testing whether the mean brain weight is the same for birds and fish, are these independent samples or paired observations?

b. What transformation of the brain weights improves the conditions for inference?

c. Assuming these are random samples, is there evidence to say that there is a difference in mean brain weight for the two groups?

P35. Radioactive twins. (Further details of this study are given in Chapter 4 on page 282.) Most people believe that country air is better to breathe than city air, but how would you prove it? You might start by choosing a response that narrows down what you mean by "better." One feature of healthy lungs is that they are quick to get rid of any nasty stuff they breathe in, such as particles of dust and smoke. This study used as its response variable the rate of tracheobronchial clearance, that is, how quickly the lungs got rid of nasty stuff.
Investigators managed to find seven pairs of identical twins who satisfied two requirements: (1) One twin from each pair lived in the country and the other lived in a city, and (2) both twins in the pair were willing to inhale an aerosol of radioactive Teflon particles. That was how the investigators measured tracheobronchial clearance. The level of radioactivity was measured twice for each twin—right after inhaling and then again an hour later. The response was the percentage of original radioactivity still remaining 1 hour after inhaling. The data are shown in Display 9.85.

<table>
<thead>
<tr>
<th>Environment</th>
<th>Rural</th>
<th>Urban</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twin Pair</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10.1</td>
<td>28.1</td>
</tr>
<tr>
<td>2</td>
<td>51.8</td>
<td>36.2</td>
</tr>
<tr>
<td>3</td>
<td>33.5</td>
<td>40.7</td>
</tr>
<tr>
<td>4</td>
<td>32.8</td>
<td>38.8</td>
</tr>
<tr>
<td>5</td>
<td>69.0</td>
<td>71.0</td>
</tr>
<tr>
<td>6</td>
<td>38.8</td>
<td>47.0</td>
</tr>
<tr>
<td>7</td>
<td>54.6</td>
<td>57.0</td>
</tr>
</tbody>
</table>


a. Discuss whether a statistical test is appropriate.
b. Regardless of your answer to part a, carry out the test.
c. Tell how to design a study that uses two independent samples (as an alternative to paired samples) to compare tracheobronchial clearance rates for people living in urban and rural environments.
d. What are the advantages of using paired data from a single sample? 
e. What are the advantages of using two independent samples?

Is It Meaningful to Compare Means?
P36. Hens' eggs. In the late 1960s, Harvard statistician Arthur Dempster went into his kitchen and measured the length, width, and volume of a dozen hens' eggs. Display 9.86 gives the lengths and widths, in inches. It would be possible to use these numbers to test the null hypothesis that the lengths and widths are equal. Do you think that would be a sensible comparison? Why or why not?

<table>
<thead>
<tr>
<th>Length</th>
<th>Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.15</td>
<td>1.89</td>
</tr>
<tr>
<td>2.09</td>
<td>1.86</td>
</tr>
<tr>
<td>2.10</td>
<td>1.87</td>
</tr>
<tr>
<td>2.14</td>
<td>1.87</td>
</tr>
<tr>
<td>2.19</td>
<td>1.87</td>
</tr>
<tr>
<td>2.13</td>
<td>1.85</td>
</tr>
</tbody>
</table>


P37. An educational researcher is interested in determining whether calculus students who study with music playing perform differently from those who study with no music playing.

a. She asks for volunteers to participate in the study and finds a large number of students who study with music playing and a large number who study with no music playing. She then compares the mean scores on the next calculus exam, using a *t*-test. Is this an appropriate design and analysis for the problem at hand? Explain your reasoning.

b. Realizing that the abilities and backgrounds of the students might be important in this study, the researcher pairs students (one student to each treatment group) from the volunteer groups on the basis of their current grade in the class. She then compares the mean scores on the next calculus exam, using a *t*-test. Is this an appropriate design and analysis for this problem? Is it better than the design in part a?

c. Explain how you would design a study to compare the effects of the two treatments on students’ performance in calculus class.
Exercises

E69. You would expect that teams launching gummy bears would get better with practice. Part of the bears-in-space data from Chapter 4 (page 258) is given in Display 9.87. You will test whether there is statistically significant evidence that teams improved from launch 1 to launch 10. Use $\alpha = 0.01$.

<table>
<thead>
<tr>
<th>Team</th>
<th>First Launch (in.)</th>
<th>Tenth Launch (in.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>74</td>
</tr>
<tr>
<td>2</td>
<td>44</td>
<td>120</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>92</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>33</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>37</td>
</tr>
<tr>
<td>6</td>
<td>125</td>
<td>174</td>
</tr>
</tbody>
</table>

Display 9.87  Selected bears-in-space data.

a. Should you use a two-sample test of the difference of two means or a one-sample test of the mean of the differences? Explain. Should this be a one-sided test or a two-sided test?

b. State hypotheses for your test, check conditions, do the computations, and write a conclusion in context.

c. What type of error might you have made?

d. Does the result of the test seem sensible compared with the result in E69?

E70. Refer to E69. This time you will test to see if there is statistically significant evidence that teams improved from launch 5 to launch 10. Use $\alpha = 0.01$. The data are given in Display 9.88.

<table>
<thead>
<tr>
<th>Team</th>
<th>Fifth Launch (in.)</th>
<th>Tenth Launch (in.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19</td>
<td>74</td>
</tr>
<tr>
<td>2</td>
<td>37</td>
<td>120</td>
</tr>
<tr>
<td>3</td>
<td>88</td>
<td>92</td>
</tr>
<tr>
<td>4</td>
<td>41</td>
<td>33</td>
</tr>
<tr>
<td>5</td>
<td>65</td>
<td>37</td>
</tr>
<tr>
<td>6</td>
<td>125</td>
<td>174</td>
</tr>
</tbody>
</table>

Display 9.88  Selected bears-in-space data.

a. Should you use a test of the difference of two means or a test of the mean of the differences? Explain. Should this be a one-sided test or a two-sided test?

b. State hypotheses for your test, check conditions, do the computations, and write a conclusion in context.

c. What type of error might you have made?

d. Does the result of the test seem sensible compared with the result in E69?

E71. Display 9.89 gives the average distance soared by the gummy bears launched by the six teams under two conditions: one book under the ramp and four books under the ramp. This was a randomized block design, with a team being the block. Whether the team did one book first or four books first was decided by the flip of a coin. You want to estimate the mean difference of the distance soared (four books — one book) with 95% confidence.

<table>
<thead>
<tr>
<th>Team</th>
<th>One Book (distance in inches)</th>
<th>Four Books (distance in inches)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>87</td>
<td>246</td>
</tr>
<tr>
<td>2</td>
<td>43</td>
<td>67</td>
</tr>
<tr>
<td>3</td>
<td>87</td>
<td>244</td>
</tr>
<tr>
<td>4</td>
<td>81</td>
<td>38</td>
</tr>
<tr>
<td>5</td>
<td>49</td>
<td>103</td>
</tr>
<tr>
<td>6</td>
<td>44</td>
<td>64</td>
</tr>
</tbody>
</table>

Display 9.89  Gummy bear launch data.

a. Should you use a one-sample procedure or a two-sample procedure? Explain.

b. Are the conditions satisfied for constructing a confidence interval using your procedure?

c. Construct and interpret the 95% confidence interval.

E72. An undesirable side effect of some antihistamines is drowsiness, which can be measured by the flicker frequency of patients (number of flicks of the eyelids per minute). Low flicker frequency is
related to drowsiness because the eyes stay shut too long. One study reported data for nine subjects (Display 9.90), each given meclastine (A), a placebo (B), and promethazine (C), in random order. At the time of the study, A was a new drug and C was a standard drug known to cause drowsiness.

<table>
<thead>
<tr>
<th>Patient</th>
<th>Drug A</th>
<th>Drug B</th>
<th>Drug C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>31.25</td>
<td>33.12</td>
<td>31.25</td>
</tr>
<tr>
<td>2</td>
<td>26.63</td>
<td>26.00</td>
<td>25.87</td>
</tr>
<tr>
<td>3</td>
<td>24.87</td>
<td>26.13</td>
<td>25.87</td>
</tr>
<tr>
<td>4</td>
<td>28.75</td>
<td>29.63</td>
<td>29.87</td>
</tr>
<tr>
<td>5</td>
<td>28.63</td>
<td>28.37</td>
<td>24.50</td>
</tr>
<tr>
<td>6</td>
<td>30.63</td>
<td>31.25</td>
<td>29.37</td>
</tr>
<tr>
<td>7</td>
<td>24.00</td>
<td>25.50</td>
<td>23.87</td>
</tr>
<tr>
<td>8</td>
<td>30.12</td>
<td>28.50</td>
<td>27.87</td>
</tr>
<tr>
<td>9</td>
<td>25.13</td>
<td>27.00</td>
<td>24.63</td>
</tr>
</tbody>
</table>


a. Are the conditions met for doing inference on the mean difference between drugs A and B? Between drugs B and C? Between drugs A and C?

b. Using a 95% confidence interval, estimate the mean of the differences between drugs A and B. Save your conclusion for part e.

c. Using a 95% confidence interval, estimate the mean of the differences between drugs B and C. Save your conclusion for part e.

d. Using a 95% confidence interval, estimate the mean of the differences between drugs A and C. Save your conclusion for part e.

e. Write a summary of what you learned about the three treatments in this analysis.

E73. Suppose the aldrin data (see page 620) actually had been collected so that the data were paired. That is, the bottom and mid-depth measurements in the same row in Display 9.91 were taken at the same time and in the same place.

<table>
<thead>
<tr>
<th>Bottom</th>
<th>Mid-Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.8</td>
<td>3.2</td>
</tr>
<tr>
<td>4.8</td>
<td>3.8</td>
</tr>
<tr>
<td>4.9</td>
<td>4.3</td>
</tr>
<tr>
<td>5.3</td>
<td>4.8</td>
</tr>
<tr>
<td>5.4</td>
<td>4.9</td>
</tr>
<tr>
<td>5.7</td>
<td>5.2</td>
</tr>
<tr>
<td>6.3</td>
<td>5.2</td>
</tr>
<tr>
<td>7.3</td>
<td>6.2</td>
</tr>
<tr>
<td>8.1</td>
<td>6.3</td>
</tr>
<tr>
<td>8.8</td>
<td>6.6</td>
</tr>
</tbody>
</table>

Display 9.91  Paired aldrin data.

a. Construct a 95% confidence interval for the mean difference in the aldrin concentrations. Do you have statistically significant evidence that the mean difference is not 0?

b. Do you reach the same conclusion as in the example on page 626?

c. Suppose the observations are paired and so are not independent and the null hypothesis is false. Which gives more power, a one-sample test of the mean difference or a two-sample test of the difference of the means?

E74. In the fitting of hearing aids to individuals, it is standard practice to test whether a particular hearing aid is right for a patient by playing a tape on which 25 words are pronounced clearly and then asking the patient to repeat the words as heard. Different lists are used for this purpose, and, in order to make accurate checks of hearing, the lists are supposed to be of equal difficulty with regard to understanding them correctly. The research question of interest is “Are the test lists equally difficult to understand in the presence of background noise?” Display 9.92 shows the number of words identified correctly out of a list of 50 words by 24 people with normal hearing, in the presence of background noise. The
two different lists were presented to each person in random order.

a. First, suppose the data are not paired. For example, in the first row, the 28 words in list 1 and the 20 words in list 2 are not for the same person. Use the information in Display 9.93 to decide whether the conditions appear to be met for conducting a two-sample test to determine whether the mean number of words understood differs between the two lists.

Display 9.92  Number of correctly identified words in two lists. [Source: lib.stat.cmu.edu.]

b. Do the computations for the two-sample t-test. Record the P-value.

c. Now use the fact that each row consists of the two measurements for the same person. Are the conditions met for conducting a test of significance for the mean difference?

d. Do the computations for the one-sample t-test. Record the P-value.

e. Which test is more powerful, the two-sample t-test or the one-sample t-test?

E75. Aerobic exercise is good for the heart. A group of college students interested in the effect of stepping exercises on heart rate conducted an experiment in which subjects were randomly assigned to a stepping exercise on either a 5.75-in. step (coded 0) or an 11.50-in. step (coded 1). Each subject started with a resting heart rate and performed the exercise for 3 minutes, at which time his or her exercise heart rate was recorded. (This is a simplification of the actual design.) The data are shown in Display 9.94.

Display 9.93  Boxplots for word list data.

Display 9.94  Resting and exercise heart rates (in beats per minute), categorized by height of step. [Source: lib.stat.cmu.edu.]
a. You want to test whether the higher step resulted in a larger mean gain in heart rate than did the lower step. Is this a one-sample test or a two-sample test? Explain.

b. Check conditions and then test whether the higher step resulted in a larger mean increase in heart rate.

e76. Refer to e75. Construct and interpret a 95% confidence interval to help you decide whether the initial random assignment of treatments to subjects did a satisfactory job of equalizing the mean resting heart rates between the two treatment groups.

e77. Bee stings. Beekeepers sometimes use smoke from burning cardboard to reduce their risk of getting stung. It seems to work, and you might suppose the smoke acts like an insect repellent—but that’s not the case. When J. B. Free, at Rothamsted Experimental Station, jerked a set of muslin-wrapped cotton balls in front of a hive of angry bees, the number of stingers left by the bees made it quite clear they were just as ready to sting smoke-treated balls as they were to sting the untreated controls. Yet in tests using control cotton balls and cotton balls treated with the repellent citronellol, the bees avoided the repellent. On 70 trials out of 80, the control cotton balls gathered more stings. What, then, is the effect of the smoke?

One hypothesis is that smoke masks some other odor that induces bees to sting; in particular, smoke might mask some odor a bee leaves behind along with its stinger when it drills its target, an odor that tells other bees “Sting here.” To test this hypothesis, Free suspended 16 cotton balls on threads from a square wooden board, in a four-by-four arrangement. Eight cotton balls had been freshly stung; the other 8 were pristine and served as controls. Free jerked the square up and down over a hive that was open at the top and then counted the number of new stingers left behind in the treated and the control cotton balls. He repeated this whole procedure eight more times, each time starting with 8 fresh and 8 previously stung cotton balls and counting the number of new stingers left behind. For each of the nine occasions, he lumped together the 8 balls of each kind and took as his response variable the total number of new stingers. His data are shown in display 9.95.

<table>
<thead>
<tr>
<th>Cotton Balls</th>
<th>Occasion</th>
<th>Previously Stung</th>
<th>Fresh</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>33</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>33</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>22</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>21</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>33</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>70</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>


a. The research hypothesis was that the previously stung cotton balls would receive more stings. State the null and alternative hypotheses for a one-sample test of significance of the mean difference.

b. Are the conditions met for doing a test of the significance of the mean difference? If not, how do you suggest that you proceed?

c. Conduct your analysis and give your conclusion.
E78. In the experiment described in the leprosy bacilli example on page 648, a control group received a treatment that had no medical value. Which subjects were in the antiseptic group and which were in the control group was determined by random assignment. The results for the control group are given in Display 9.96.

<table>
<thead>
<tr>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>13</td>
</tr>
<tr>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>21</td>
<td>23</td>
</tr>
<tr>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>12</td>
<td>16</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
</tr>
</tbody>
</table>

Display 9.96  Bacteria counts before and after control treatment.

a. Does a square root transformation improve conditions for inference?
b. Construct a 95% confidence interval to determine if the control “treatment” is effective in reducing bacteria counts.

E79. Hospital carpets. If you had to spend time in a hospital, would you want your room to have carpeting or a bare tile floor? Carpeting would keep down the noise level, which certainly would make the atmosphere more restful, but bare floors might be easier to keep free of germs. One way to measure the bacteria level in a room would be to pump air from the room over a growth medium, incubate the medium, and count the number of colonies of bacteria per cubic foot of air. This method was in fact used to compare the bacteria levels in 16 rooms at a Montana hospital. Eight randomly chosen rooms had carpet installed; the floors in the other eight rooms were left bare. The data are shown in Display 9.97.

<table>
<thead>
<tr>
<th>Carpeted Floors</th>
<th>Bare Floors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Room</td>
<td>Colonies/ft³</td>
</tr>
<tr>
<td>212</td>
<td>11.8</td>
</tr>
<tr>
<td>216</td>
<td>8.2</td>
</tr>
<tr>
<td>220</td>
<td>7.1</td>
</tr>
<tr>
<td>223</td>
<td>13.0</td>
</tr>
<tr>
<td>225</td>
<td>10.8</td>
</tr>
<tr>
<td>226</td>
<td>10.1</td>
</tr>
<tr>
<td>227</td>
<td>14.6</td>
</tr>
<tr>
<td>228</td>
<td>14.0</td>
</tr>
</tbody>
</table>


b. Tell how to run the experiment using the other design.

E80. Hospital carpets. Conduct the test that is appropriate for the design you selected in part a of process described in E79.

E81. Too tight? You might have lived your entire life (until now) without once wondering what statistics has to do with the screw cap on a bottle of hair conditioner. (Brace yourself!) The machine that puts on the cap must apply just the right amount of turning force—too little, and conditioner can leak out; too much, and the cap might be damaged and be hard to get off. Imagine that you are in charge of quality control for a hair conditioner manufacturer. An engineer suggests that new settings on the capping machine will reduce the variability in the amount of force applied to the bottle caps so that fewer will be either too tight or too loose and more will be in the acceptable range. He offers to set up a comparison: Ten batches of bottles will be capped by each process, and the force needed to unscrew the cap will be measured. To make the comparison scientific, statistical methods will be used to test

\[ H_0: \mu_{new} = \mu_{old} \]  versus  \[ H_a: \mu_{new} > \mu_{old} \]

where \( \mu \) in each case is the underlying true mean for the amount of force. What do you say to the engineer?
E82. One of the important measures of quality for a product that consists of a mixture of ingredients, such as cake mix, lawn fertilizer, or powdered medications, is how well the components are mixed. This is tested by taking small samples of the product from various places in the production process and measuring the proportions of various components in those samples. The sample data then will be a set of proportions for component A, a set of proportions for component B, and so on. Does a comparison of the means of these sets of proportions help measure the degree of mixing? If not, suggest another way to assess the degree of mixing.

E83. *Guessing distances.* Imagine guessing the width of your classroom. Do you think you’d be as accurate guessing distances in meters as in feet? Not long after the metric system was officially introduced into Australia, a college professor asked a group of students to guess the width of their lecture hall to the nearest meter. He asked a different group to guess the width to the nearest foot. Display 9.98 presents summary data for their guesses.

<table>
<thead>
<tr>
<th>Estimate in Feet</th>
<th>Estimate in Meters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td>69</td>
</tr>
<tr>
<td>Sample Mean</td>
<td>43.68</td>
</tr>
<tr>
<td>Sample SD</td>
<td>12.51</td>
</tr>
</tbody>
</table>

Display 9.98  Summary data for guessing the width of a lecture hall.  [Source: M. Hills and the M345 course team, *M345 Statistical Methods, Unit 1: Data, Distributions, and Uncertainty*, Milton Keynes, The Open University (1986).]

a. Do the numbers come from a single sample of paired values, or from two independent samples?

b. Tell how to gather data in the other format.

c. Why do you think the study was done the way it was instead of the way you described in part b?

d. A test of the null hypothesis that the two population means are equal rejects that hypothesis with a $P$-value less than 0.0001. Is this a meaningful comparison? Why or why not?

E84. *Buying jeans.* Would you guess that people tend to buy more jeans when the weather is warm or when it is cold? Display 9.99 shows the total sales, in thousands of pairs, for jeans sold in the United Kingdom in January and in June for the years 1980–85. A test of the null hypothesis $H_0: \mu_{\text{Jan}} = \mu_{\text{Jun}}$ gives a $P$-value of 0.06. Is this a reasonable use of significance testing?

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>1998</td>
<td>1924</td>
<td>1969</td>
<td>2149</td>
<td>2319</td>
<td>2137</td>
</tr>
<tr>
<td>June</td>
<td>2286</td>
<td>1979</td>
<td>2375</td>
<td>2369</td>
<td>3126</td>
<td>2117</td>
</tr>
</tbody>
</table>


E85. *Old Faithful.* The Old Faithful geyser, in Yellowstone National Park, got its name from the predictability of its eruptions, which used to occur about every 66 minutes. For decades, visitors throughout the warm months of the year have crowded the large circle of benches surrounding the geyser as each eruption time approaches, waiting to see a 150-ft-high spout of superheated steam and water shoot into the air. In recent years, however, scientists have noticed that the times between eruptions have become somewhat more variable, and the distribution of times in fact appears to be bimodal, as seen in Display 9.100. Additional study suggests that the time until the next eruption might depend on the duration of the previous eruption. To test this hypothesis, a science class decides to classify eruptions as long or short and then compare the mean time until the next eruption for long eruptions and for short eruptions.
9.5 Paired Comparisons


a. Is this a meaningful comparison? Why or why not?

b. Are the data on times between eruptions paired or unpaired?

E86. Old Faithful revisited. Suppose you have the data in Display 9.100 on times between eruptions but you don't have the data on the corresponding durations. "Aha!" you think. "We want to know whether the difference between the two modes is a 'real' pattern, that is, too big to be due just to chance variation. To test this, I'll just divide the histogram in half at the low point between the two modes and compare the means of the upper and lower halves of the data using a statistical test." Is this a reasonable application of hypothesis testing?

E87. Auditors often are required to use sampling to check various aspects of accounts they are auditing. For example, suppose the accounts receivable held by a firm are being audited. Auditors typically will sample some of these accounts and compare the value the firm has on its books (book value) with what they agree is the correct amount (audit value), usually corroborated by talking with the customer. (Obtaining the audit values is a fair amount of work, hence the need for sampling.) Data of this sort are privileged information in most firms, but Display 9.101 (on the next page) shows what a typical data set might look like. Note that many of the book values and audit values agree, as they should. There is no question about the fact that these are paired data, so an estimate of the mean difference between the book values and audit values should proceed from the differences. The question is what to do with the differences that equal 0.

a. Should the estimate of the mean difference between book values and audit values be based on the entire sample of 40, or on only the 15 nonzero values? Answer this question by working through these steps.

i. Construct a 95% confidence interval estimate using all 40 differences.

ii. Construct a 95% confidence interval estimate using only the 15 nonzero differences.

iii. How do the intervals in parts i and ii compare? Which do you think provides the better answer? Why?

b. If the estimate of the mean difference is based on the nonzero values in the sample, explain how you would use the result to estimate the total amount by which the book values exceed the audit values for all accounts in the population. The firm knows how many accounts receivable it has; assume this number is 1000. (Keep in mind that the total could be negative if audit values tend to exceed book values.)
E88. In E7 on page 578, you read about the readability of cancer brochures. Can patients read the information they are given? Based on suspicions that the reading level of patient brochures is higher than the reading level of patients, a study was conducted to compare the two. A page was randomly selected from each of 30 brochures for cancer patients published by the American Cancer Society and the National Cancer Institute. The pages were judged for readability using standard readability tests. Then 63 patients were tested for their reading level. The results are summarized in Display 9.102.

<table>
<thead>
<tr>
<th>Reading Grade Level</th>
<th>Frequency for Brochures</th>
<th>Frequency for Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>17</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Display 9.102 Frequency of various reading levels, as measured for brochures and patients. [Source: Thomas H. Short et al., “Readability of Educational Materials for Patients with Cancer,” Journal of Statistics Education 3, no. 2 (July 1995).]

a. Is there evidence that the materials are at a higher mean reading level than the patients’ mean reading level? To answer this question, should you do a one-sample test or a two-sample test?

b. Perform the test you selected in part a and state your conclusion.

c. The question in part a isn’t really the important one. Answer this more
important question: If a patient reads at the 10th-grade level and is given one of the booklets at random, what is the probability that the booklet will be at a higher reading level than that of the patient?

d. If each of these patients is given one booklet at random, how many patients do you expect to get a booklet that is at too high a reading level for them?

Chapter Summary

To use the confidence intervals and significance tests of this chapter, you need either

- a random sample from a population (consisting of either single values or paired values)
- two independent random samples from two distinct populations or, in the case of an experiment, two treatments randomly assigned to the available subjects

When there is no randomness involved, you proceed with the test only after stating loudly and clearly the limitations of what you are doing. If you reject the null hypothesis in such a case, all you can conclude is that something happened that can't reasonably be attributed to chance.

You use a confidence interval if you want to find a range of plausible values for

- $\mu$, the mean of your single population
- $\mu_1 - \mu_2$, the difference between the means of your two populations

Both of the confidence intervals you studied have the same form:

$$\text{statistic} \pm (\text{critical value}) \cdot (\text{standard deviation of statistic})$$

The degrees of freedom for the two-sample $t$-procedure must be approximated from a rather cumbersome formula. For that reason, the two-sample $t$-procedures should be done with the aid of a calculator or statistics software.

All the significance tests you have studied include the same steps:

1. Justify your reasons for choosing this particular test. Discuss whether the conditions are met and, if they are not strictly met, decide whether it is okay to proceed.
2. State the null hypothesis and the alternative hypothesis.
3. Compute the test statistic, $\frac{\text{statistic} - \text{parameter}}{\text{standard deviation of statistic}}$, and find the $P$-value. Draw a sketch of the situation.
4. Use the computations to decide whether to reject or not reject the null hypothesis by comparing the $P$-value to the level of significance, $\alpha$. Then state your conclusion in terms of the context of the situation. (“Reject $H_0$” is not sufficient.) Mention any doubts you have about the validity of that conclusion.
There's only one way you can be confident of drawing correct conclusions from data: Use sound methods of data production, either random samples or randomized experiments. At the other extreme, there are many ways you can be confident that the conclusions might not be valid. For example, voluntary response samples are worthless when it comes to inference. In reality, many situations fall between the two extremes. What then? For example, what if you use your class as a sample instead of taking a random sample? There are no formal rules; the value of the inference methods is rarely all-or-nothing. The more reasonable it is to regard your data as coming from a random sample or a randomized experiment, the more reasonable it is to trust conclusions based on the inference methods. For many data sets, making a careful judgment about this issue is the hardest aspect of a statistician's job.

**Review Exercises**

E89. To find an estimate of the number of hours that highly trained athletes sleep each night, a researcher selects a random sample of 15 highly trained athletes and asks each how many hours of sleep he or she gets each night. The results are given in Display 9.103.

Display 9.103  Hours of sleep for highly trained athletes.

<table>
<thead>
<tr>
<th>Leaf Unit = 0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 7</td>
</tr>
<tr>
<td>4 29</td>
</tr>
<tr>
<td>5 149</td>
</tr>
<tr>
<td>6 15</td>
</tr>
<tr>
<td>7 115</td>
</tr>
<tr>
<td>8 58</td>
</tr>
<tr>
<td>9 39</td>
</tr>
</tbody>
</table>

a. Are the conditions satisfied for computing a confidence interval for the mean?

b. Construct and interpret a 95% confidence interval for the mean.

E90. For the sleep data in E89, carry out a test of the claim that the population mean is 8 hours of sleep per night.

a. State the hypotheses, defining any symbols that you use.

b. Calculate the value of the test statistic, showing the formula and the values substituted in.

c. Find the P-value from your calculator and interpret it in the context of this situation.

E91. Seventy students were randomly assigned to launch a gummy bear, by themselves, using either one book or four books (see page 258). Display 9.104 gives the results of all 70 first launches by these students. Is there statistically significant evidence of a difference in the mean distance? Show all four steps of a test of significance, being sure to name the test you are using.

Display 9.104  Summary information for gummy bear launches.
E92. *Old Faithful.* (Refer to E85 on page 660 for a description of the data.) Display 9.105 presents summary statistics and side-by-side boxplots comparing the distributions of the times until the next eruption following short eruptions (duration \(< 3\) min) and long eruptions (duration \(\geq 3\) min).

<table>
<thead>
<tr>
<th></th>
<th>Short Eruptions (min) (n = 67)</th>
<th>Long Eruptions (min) (n = 155)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean</strong></td>
<td>54.46</td>
<td>78.16</td>
</tr>
<tr>
<td><strong>SD</strong></td>
<td>6.30</td>
<td>6.89</td>
</tr>
</tbody>
</table>

Display 9.105  Summary statistics and boxplots of times until next eruption following short and long eruptions.

a. State appropriate null and alternative hypotheses in words. Then define notation and restate \(H_0\) and \(H_A\) in symbols.
b. Tell whether the design of the study justifies the use of a probability model, and give reasons for your answer.
c. Based on the summary statistics and boxplots, tell whether the shapes of the distributions raise doubts about using a \(t\)-test.
d. Carry out the computations, find the \(P\)-value, and tell what your conclusion is if you take the results at face value.
e. Now tell what you think the results of the test really mean for this particular data set.

E93. The data in Display 9.106 show the life expectancies (in years) of males and females for a random sample of African countries. Is there evidence of a significant difference between the life expectancies of males and females in the countries of Africa? Give statistical justification for your answer and a careful explanation of your analysis.

<table>
<thead>
<tr>
<th>Country</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algeria</td>
<td>71.5</td>
<td>74.6</td>
</tr>
<tr>
<td>Gambia</td>
<td>51.9</td>
<td>55.6</td>
</tr>
<tr>
<td>Ghana</td>
<td>57.7</td>
<td>59.3</td>
</tr>
<tr>
<td>Kenya</td>
<td>48.9</td>
<td>47.1</td>
</tr>
<tr>
<td>Liberia</td>
<td>37.0</td>
<td>40.8</td>
</tr>
<tr>
<td>Libya</td>
<td>74.3</td>
<td>78.8</td>
</tr>
<tr>
<td>Mauritius</td>
<td>68.4</td>
<td>76.4</td>
</tr>
<tr>
<td>Morocco</td>
<td>68.4</td>
<td>73.1</td>
</tr>
<tr>
<td>Swaziland</td>
<td>32.5</td>
<td>34.0</td>
</tr>
<tr>
<td>Togo</td>
<td>55.0</td>
<td>59.1</td>
</tr>
</tbody>
</table>

Display 9.106  Life expectancies (in years) for a sample of African countries. [Source: CIA Factbook, 2006.]

E94. *Altitude and alcohol.* On every commercial passenger flight, there is an announcement that tells what to do if the cabin loses pressure. At high altitudes, air has less oxygen, and you can lose consciousness in a short time. In 1965, the *Journal of the American Medical Association* published a paper reporting the effects of alcohol on the length of time subjects stayed conscious at high altitudes.

There were ten subjects. Each was put in an environment equivalent to an altitude of 25,000 ft and then monitored to see how long the subject could continue to perform a set of assigned tasks. As soon as performance deteriorated (the end of “useful consciousness”), the time was recorded and the environment was returned to normal.

Three days later, each subject drank a dose of whiskey based on body weight—1 cc of 100-proof alcohol for every 2 lb—and then, after waiting 1 hour for the whiskey to take effect, was returned to the simulated altitude of 25,000 ft for another test. Display 9.107 gives the times (in seconds) until the end of useful consciousness under both conditions.

a. How can you tell from simply the description of the data, even before seeing the numbers, that this study gives you paired data from one sample rather than two independent samples?

b. Carry out an exploratory analysis. Include stemplots or boxplots of the recorded times under each of the two conditions and a stemplot or boxplot of the differences, as well as a scatterplot. Describe the patterns in words, and tell what questions, if any, the patterns raise about the validity of formal inference based on the normal distribution.

c. Find a 95% confidence interval for the “true” difference, \( \mu_C - \mu_A \). What does the “true” difference refer to in this context? What do you conclude about the effect of alcohol on the time of useful consciousness at high altitude?

E95. Select the best answer. The confidence level measures
A. the fraction of times the confidence interval will capture the parameter it is estimating in repeated use of the procedure on different samples
B. the fraction of times the confidence interval will fail to capture the parameter it is estimating in repeated use of the procedure on different samples
C. the fraction of times the confidence interval captures the sample statistic on which it is based
D. None of the above is a good interpretation.

E96. Select the best answer. The confidence level measures
A. the fraction of times the confidence interval will capture the parameter it is estimating in repeated use of the procedure on different samples
B. the fraction of times the confidence interval will fail to capture the parameter it is estimating in repeated use of the procedure on different samples
C. the fraction of times the confidence interval captures the sample statistic on which it is based
D. None of the above is a good interpretation.

E97. The distribution of annual incomes of the employees of a large firm is highly skewed toward the larger values because of the high salaries of the upper-level managers. A random sample of size \( n \) is to be selected for the purpose of testing the claim that the mean salary is over $50,000.

a. How would you proceed if \( n = 10 \)?

b. How would you proceed if \( n = 50 \)?

E98. Boosting your SATs? Several commercial companies offer special courses designed to help you get a higher score on the SAT. Suppose market research shows that students are willing to take such a course only if, on average, it raises SAT scores by at least 30 points. Here’s a way to test a claim that a course is able to do that. You’ll need some volunteers, the more representative the better. Randomly divide them into two groups. Those in the first group take the special course; those in the second group serve as controls. When the course is over, both groups take the SAT.

Let \( \bar{x}_1 \) and \( \bar{x}_2 \) be the sample means for the two groups, let \( s_1 \) and \( s_2 \) be the sample standard deviations, and let \( n_1 \) and \( n_2 \) be the sample sizes.

a. Define suitable notation, and state the null and alternative hypotheses twice, first in words and then in symbols.
b. If you construct a 95% confidence interval for $\mu_1 - \mu_2$, under what circumstances do you reject $H_0$?

c. Carry out the test using the $t$-statistic for $\bar{x}_1 = 1100, \bar{x}_2 = 1060, s_1 = 100, s_2 = 80$, and $n_1 = n_2 = 16$. Check your conclusion by constructing a confidence interval.

E99. If you consider all regular players (not counting pitchers) on Major League Baseball teams over recent years, the season batting averages are approximately normally distributed, with a mean around .250 and a standard deviation around .050. Carefully explain each component of a fixed-level test in the context of these questions. Use $\alpha = 0.05$.

a. The New York Yankees had a team batting average of .276 in 2005, with 16 regular players contributing. Is this a higher average than would be expected for a random sample of 16 players?

b. The Seattle Mariners had a team batting average of .256 in 2005, also with about 16 regular players contributing. Can this team be considered above average in hitting? [Source: www.mlb.com, October 2005.]

E100. Carbon monoxide. The data in Display 9.108 give the “yield” of carbon monoxide (CO) for sidestream smoke (what the smoker doesn’t inhale) and mainstream smoke (what the smoker inhales) from eight brands of Canadian cigarettes.

<table>
<thead>
<tr>
<th>Cigarette Brand</th>
<th>Sidestream (mg)</th>
<th>Mainstream (mg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>40.5</td>
<td>18.6</td>
</tr>
<tr>
<td>B</td>
<td>59.8</td>
<td>20.5</td>
</tr>
<tr>
<td>C</td>
<td>42.9</td>
<td>16.8</td>
</tr>
<tr>
<td>D</td>
<td>42.0</td>
<td>17.8</td>
</tr>
<tr>
<td>E</td>
<td>60.8</td>
<td>19.8</td>
</tr>
<tr>
<td>F</td>
<td>45.1</td>
<td>16.4</td>
</tr>
<tr>
<td>G</td>
<td>43.9</td>
<td>13.1</td>
</tr>
<tr>
<td>H</td>
<td>67.3</td>
<td>12.4</td>
</tr>
</tbody>
</table>

AP1. In measuring the angle formed by two intersecting laser beams in a physics lab, a student uses chalk dust to illuminate the beams and then uses a protractor to measure the angle between them. She takes ten measurements and then produces a 95% confidence interval estimate of the mean, but is unhappy with the large margin of error that she calculates. Which of the following is the worst plan for reducing the margin of error the next time she takes similar measurements?

A. Increase the number of measurements.
B. Use a more precise method of measuring angles.
C. Decrease the confidence level.
D. Check to be sure any outliers aren’t mistakes.
E. Combine her results with those from other students.

AP2. You plan to take a random sample of 20 cars from your neighborhood to see whether it’s reasonable to assume that the mean age of all cars in your neighborhood is 7 years. You know that the population of all the ages of cars in your neighborhood is likely to be strongly skewed to the right. Which of the following is not a good reason for doing a log transformation on your data?

A. Because the population is likely to be strongly skewed to the right
B. To make the distribution of the values more symmetric
C. To decrease the probability of a Type I error
D. To decrease the value of the mean
E. To help ensure that the probability that a 95% confidence interval will capture the population mean is 0.95

AP3. Which of the following describes a difference between the sampling distribution of
\[ z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \]
and the sampling distribution of
\[ t = \frac{\bar{x} - \mu}{s / \sqrt{n}} \]
in the case where \( n = 5 \)?

A. The distribution of \( z \) is more symmetric.
B. The distribution of \( t \) is more skewed.
C. The distribution of \( t \) has a larger spread.
D. The distributions have different means.
E. The distribution of \( t \) is known, while the distribution of \( z \) is unknown because \( \sigma \) is unknown.

AP4. With a sample of size 15 and \( t = -2.76 \), should the null hypothesis be rejected in a two-sided significance test for a mean?

A. Yes if \( \alpha = 0.05 \), and yes if \( \alpha = 0.01 \)
B. Yes if \( \alpha = 0.05 \), but no if \( \alpha = 0.01 \)
C. No if \( \alpha = 0.05 \), but yes if \( \alpha = 0.01 \)
D. No if \( \alpha = 0.05 \), and no if \( \alpha = 0.01 \)
E. Never, because \( t \) is negative

AP5. Researchers wish to estimate, for the population of married couples, the average of the differences in the heart rate of each wife and her husband. They get a random sample of 100 married couples and measure the heart rate of each person. What is the best way to proceed?

A. Get a random sample of unmarried couples for comparison.
B. The independence assumption has been violated, so the researchers must start over and get a random sample of married women and, independently, get a random sample of married men.
C. Compute the mean heart rate of the women and the mean heart rate of the men. Compute a confidence interval for the difference between the mean heart rate of the women and the mean heart rate of the men.
Compute the mean heart rate of the women and the mean heart rate of the men. Compute separate confidence intervals for the mean heart rate of the women and for the mean heart rate of the men.

Subtract each husband’s heart rate from his wife’s heart rate, find the mean of these differences, and compute a single confidence interval for the mean difference.

AP6. To estimate the mean number of hours that students at a particular college sleep on a school night, a dean selects a random sample of 16 students and asks each how many hours they sleep on a school night. The number of hours are given in the stem-and-leaf plot below. The dean checks, and the outlier isn’t a mistake. What is the best way for the dean to proceed?

<table>
<thead>
<tr>
<th>Stem-and-leaf of Number of Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>N = 16</td>
</tr>
<tr>
<td>Leaf Unit = 0.10</td>
</tr>
<tr>
<td>5 9</td>
</tr>
<tr>
<td>6 1 2 5 7</td>
</tr>
<tr>
<td>7 5 9 9</td>
</tr>
<tr>
<td>8 3 5 8 9</td>
</tr>
<tr>
<td>9 1 2 6</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>12 9</td>
</tr>
</tbody>
</table>

Remove the outlier from the data set. This student is not at all typical.

Because it is so easy to get this information, get a larger random sample.

Do a log transformation.

Do a reciprocal transformation.

Do a square root transformation.

AP7. Scotland recently imposed a ban on smoking in bars. Before the ban, a researcher thought that the respiratory health of bar employees should improve after working in smoke-free air. Before the ban went into effect, he scored the respiratory health of a random sample of Scottish bar employees. Two months after the ban, he obtained an independent random sample of bar employees and scored their respiratory health. The increase in the mean scores was statistically significant ($P = 0.049$). Which of the following is the best interpretation of this result?

A An observed difference in sample means as large or larger than that in this sample is unlikely to occur if the mean score for all bar employees before the ban is equal to the mean after the ban.

B The probability is only 0.049 that the mean score for all bar employees increased from before the ban to after the ban.

C The mean score for all bar employees increased by more than 4.9%.

D There is a 4.9% chance that the mean score of all bar employees after the ban is actually lower than before the ban, despite the increase observed in the samples.

E Only 4.9% of bar employees had their scores drop while the other 95.1% had their scores increase.

AP8. Several statistically trained knights seek to estimate the mean tail length of adult fire-breathing dragons. They collect a random sample of 10 dragons, observe that the distribution of tail lengths is symmetric with no outliers, and then compute a confidence interval. They publish a paper stating, “We are 95% confident that the mean tail length of adult fire-breathing dragons is between 8.2 feet and 11.3 feet.” The knights later discover that in the population of all dragons, there are some with exceptionally short tails due to injuries, though none of these dragons were
which of the following is the best description of how the knights should revise the interpretation of their confidence interval?

- They should lower the center.
- They should decrease the width.
- They should increase the width.
- They should decrease the confidence level.
- They should replace “adult” with “uninjured adult.”

**Investigative Tasks**


![Stereogram with embedded diamond](image)

**Display 9.109** Stereogram with embedded diamond.

In an experiment, one group (NV) of subjects was either told nothing or told that a diamond was embedded. A second group (VV) of subjects was shown a drawing of the diamond. The times (in seconds) that it took the subjects to see the diamond are summarized in Display 9.110.

![Boxplot and summary statistics](image)

**Display 9.110** Boxplot and summary statistics of time, in seconds, to see the diamond.

- a. Check the conditions for doing a two-sample *t*-test of the difference in means.
- b. The researchers did a two-sample *t*-test on the original data, with pooled variances (\( \alpha = 0.05 \)). Do you agree with that decision? Why or why not?
- c. Replicate the test that the researchers did, and find the *P*-value. What conclusion do you come to if you take the results at face value?
- d. What would the test decision have been if the variances weren’t pooled?
- e. When the standard deviations aren’t comparable, what is the effect of using the pooled procedure rather than the unpooled procedure?
AP10. Refer to the experiment described in AP9. The results of a log transformation of the times are given in Display 9.111.

a. Check the conditions for doing a two-sample \( t \)-test on the log-transformed data of the difference in means.

b. Do a two-sample \( t \)-test on the transformed data, with pooled variances. Record the \( P \)-value.

c. Do a two-sample \( t \)-test on the transformed data, with unpooled variances. Record the \( P \)-value.

d. When the standard deviations are comparable, what is the effect of using the pooled procedure rather than the unpooled procedure?
Is it necessary to have a child at some point in your life in order to feel fulfilled? A poll asked this question of samples of adults in various countries. You can use a chi-square test to decide if the distribution of answers (yes, no, and undecided) reasonably could be the same for the adults in each country.
You have now completed two chapters on statistical inference. Chapter 8 considered inference for proportions constructed from categorical variables, and Chapter 9 considered inference for means constructed from measurement (quantitative) variables. Chapter 10 returns to categorical variables. You can think of it as a generalization of the results of Chapter 8.

In Chapter 8, there were only two categories—success or failure, heads or tails, and so on—and you used significance tests to answer questions such as these:

*Fair coin?* You spin a coin 100 times and get tails 64 times. Do you have evidence that spinning the coin is unfair?

*Who watches the Super Bowl?* In a random sample of 100 college graduates and 100 high school graduates, 40% of the college graduates and 49% of the high school graduates watched the last Super Bowl. Is this convincing evidence that the percentage of college graduates and the percentage of high school graduates who watch the Super Bowl are not the same?

In this chapter, you will use chi-square tests to answer questions that are similar to these but usually involve more than two categories:

*Fair die?* You roll a die 60 times. You get 12 ones, 9 twos, 10 threes, 6 fours, 11 fives, and 12 sixes. Do you have evidence that the die is unfair?

*Super Bowl again.* In addition to the random samples of college and high school graduates, you have a random sample of 100 people who didn’t graduate from high school. Of this sample, 37% watched the last Super Bowl. Do the percentages of people watching the Super Bowl vary with educational level?

You will use the chi-square technique to answer a new type of problem as well:

*Independence of Favorites?* Each student in a random sample is asked to name his or her favorite sport and favorite type of music. Do you have evidence of an association between favorite sport and favorite type of music, or do these two variables appear to be independent?

**In this chapter, you will learn about three chi-square tests:**

- *Goodness of fit.* Are the proportions of the different outcomes in this population equal to the hypothesized proportions?

- *Homogeneity of proportions.* Are the proportions of the different outcomes in one population equal to those in another population?

- *Independence.* Are two different variables independent in this population?

The procedure in each case is almost the same. The difference is in the type of question asked and the kinds of conclusions that you can draw.
Testing a Probability Model: The Chi-Square Goodness-of-Fit Test

The chi-square test was developed by the English statistician Karl Pearson in 1900. This was more than a small contribution to statistical methodology—it transformed the way scientists view the world of data. Science 84 (a publication of the American Association for the Advancement of Science) included the chi-square test as one of the top 20 discoveries that have changed our lives, along with the invention of the computer, the laser, plastic, and nuclear energy. “The chi-square test was a tiny event in itself, but it was the signal for a sweeping transformation in the ways we interpret our numerical world. Now there could be a standard way in which ideas were presented to policy makers and to the general public. . . . For better or worse, statistical inference has provided an entirely new style of reasoning.” [Source: Science 84 (November 1984): 69–70.]

A goodness-of-fit test determines whether it is reasonable to assume that your sample came from a population in which, for each category, the proportion of the population that falls into the category is equal to some hypothesized proportion. If the result from the sample is very different from the expected results, then you have to conclude that the hypothesized proportions are wrong. To determine how different “very different” is, you will need a test statistic and its probability distribution.

A Test Statistic

To begin a chi-square goodness-of-fit test, make a table with the first column listing the possible outcomes. In the second column, give the frequency (or count) that each outcome was observed (O). In the third column, give the frequencies you would expect (E) if the hypothesized proportions are correct.

Example: Fair Die?

Suppose you roll a die 60 times and get 12 ones, 9 twos, 10 threes, 6 fours, 11 fives, and 12 sixes. Set up a table that you can use to see if you have evidence that the die is unfair.

Solution

If a fair die is rolled 60 times, you “expect” to get each face of the die on \( \frac{1}{6} \) of the 60 rolls, or 10 times each. Display 10.1 summarizes this information.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Observed Frequency, O</th>
<th>Expected Frequency, E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>60</td>
<td>60</td>
</tr>
</tbody>
</table>

Display 10.1  Observed and expected frequencies for 60 rolls of a fair die.
You will conclude that the die is unfair if the observed frequencies are far from the expected frequencies. Are they? Here’s where you need a test statistic; that is, you need to condense the information in the table into a single number that acts as an index of how far away the observed frequencies are from the expected frequencies.

Both the squared difference \((O - E)^2\) and the relative or proportional difference \((O - E)/E\) are important in determining how “far” the observed frequency is from the expected frequency. The test statistic involves both.

### The Test Statistic

The **test statistic for chi-square tests** is

\[
\chi^2 = \sum \frac{(O - E)^2}{E}
\]

Here \(O\) is the observed frequency in each category, and \(E\) is the corresponding expected frequency.

#### Example: The Test Statistic for the Die

Compute the value of \(\chi^2\) for the fair die problem in the previous example using a table.

**Solution**

Make a table as shown in Display 10.2.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Observed Frequency, (O)</th>
<th>Expected Frequency, (E)</th>
<th>((O - E)^2/E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>10</td>
<td>2/0.4 = 0.4</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>10</td>
<td>-1/0.1 = 0.1</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>10</td>
<td>0/0 = 0</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>10</td>
<td>-4/1.6 = 0.25</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>10</td>
<td>1/0.1 = 10</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>10</td>
<td>2/0.4 = 0.5</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>60</strong></td>
<td><strong>60</strong></td>
<td><strong>2.6</strong></td>
</tr>
</tbody>
</table>

Display 10.2 Table to compute \(\chi^2\) for the fair die problem.

The value of \(\chi^2\) is the sum of the values in the last column, \((O - E)^2/E\), so \(\chi^2 = 2.6\).

### Discussion

A Test Statistic

D1. What is the null hypothesis for the fair die example? What is the alternative hypothesis?
D2. One test statistic that might be constructed for the data in the fair die example is

\[ \sum (O - E) \]

where \( O \) represents an observed frequency and \( E \) represents the corresponding expected frequency.

a. Compute the value of this test statistic for the fair die example.
b. Will your result in part a always occur? Prove your answer.
c. Is this a good test statistic?
d. You have seen two other situations in which the sum of the differences always turned out to be 0. What were those situations? What did you do in those situations?

D3. Another test statistic that might be constructed is

\[ \sum (O - E)^2 \]

a. Compute the value of this test statistic for the fair die example.
b. Display 10.3 shows the results from the rolls of two different dice. For which die does the table give stronger evidence that the die is unfair?
c. Compute and compare the values of \( \sum (O - E)^2 \) for Die A and Die B. Does this appear to be a reasonable test statistic? Explain.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Die A</th>
<th>Die B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed Frequency</td>
<td>Expected Frequency</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>10</td>
</tr>
</tbody>
</table>

Display 10.3 Results from the rolls of two different dice.

D4. Will you reject the hypothesis that the die is fair if \( \chi^2 \) is relatively large, if it is relatively small, or both? How might you determine whether a \( \chi^2 \) value of 2.6 is relatively large?

D5. What is \( \chi^2 \) for Die A and for Die B in Display 10.3? Did the larger value of \( \chi^2 \) correspond to the die that you thought seemed more unfair? What purpose does dividing by \( E \) serve in the formula for the test statistic?

D6. If you are given the observed frequencies and the hypothesized proportions, how do you find the expected frequencies?

D7. The \( \chi^2 \) statistic involves a sum of squared differences. What other statistics have you seen that involve a sum of squared differences?
The Distribution of Chi-Square

In the fair die example, you saw that $\chi^2$ turned out to be 2.6. If the observed and expected frequencies had been exactly equal, $\chi^2$ would have been 0 and you would have had no reason to doubt that the die is fair. If the observed and expected frequencies had been much farther apart than they were, $\chi^2$ would have been much larger than 2.6 and you would have had evidence that the die is not fair.

How can you assess whether a value of $\chi^2$ of 2.6 is large enough to reject the hypothesis that the die is fair? You need to see how much variation there is in the value of $\chi^2$ when a die is fair. In Activity 10.1a, you will do this.

ACTIVITY 10.1a

Generating a Chi-Square Distribution

What you'll need: dice

1. Form five groups in your class. Your group should roll fair six-sided dice until your group has a total of 60 rolls. Compute $\chi^2$ for your group's results. Get the values of $\chi^2$ from the other four groups. From just these five results, does it appear that a $\chi^2$ value of 2.6 is a reasonably likely outcome when a fair die is rolled?

2. Describe how to use your calculator to simulate 60 rolls of a fair die. [See Calculator Note 10A.]

3. Use your calculator to simulate 60 rolls of a fair die. Compute $\chi^2$ for your results. [See Calculator Note 10B.]

4. Continue until your class has 200 values of $\chi^2$, where each $\chi^2$ value is computed from 60 rolls of a fair die. Display the values in a histogram and describe the shape of the histogram.

5. Should a one-sided test or a two-sided test be used to test if a die is fair? Using the results from your simulation, estimate the $P$-value for a $\chi^2$ value of 2.6. Is a $\chi^2$ value of 2.6 a reasonably likely outcome if the die is fair?

The distributions in Display 10.4 each show 2000 values of $\chi^2$ computed from rolls of a fair die.

Display 10.4 Two histograms of 2000 values of $\chi^2$, one with $n = 1000$ (left) and one with $n = 60$ (right).
In the histogram on the left in Display 10.4, each value of $\chi^2$ was computed from the results of 1000 rolls of a fair die. The histogram on the right was constructed from the results of only 60 rolls, so it should look very much like the one your class generated in Activity 10.1a.

As you can see, the $\chi^2$ distribution does not change much with a change in sample size, as long as the sample size is large. Is there only one $\chi^2$ distribution (like the $z$-statistic), or are there many (like the $t$-statistic)? Sample size doesn't change the shape, center, or spread, but there is one more variable to check—the number of categories.

Each histogram in Display 10.5 shows a distribution of 5000 values of $\chi^2$. Each of the 5000 values was computed from the results of 60 rolls of a fair die. However, each histogram was made using a die with a different number of sides. Notice how the distribution changes as the number of categories (sides) changes.

You can use a calculator to compare different chi-square probability density functions. In these calculator screens, the numbers used in the commands are number of categories − 1 (the “degrees of freedom”), so the graphs show chi-square density functions for 4, 8, and 20 categories. [See Calculator Note 10C to learn how to graph a chi-square probability density function, which are continuous approximations to distributions like those in Display 10.5.]
The Distribution of Chi-Square

D8. Refer to Display 10.5.

a. Describe how the distribution of $\chi^2$ changes as the number of categories increases. Discuss these changes in terms of shape, center, and spread.

b. For which die are you most likely to get a value of $\chi^2$ of 15 or more? For which die are you least likely to do so?

c. For each die, make a rough approximation of the smallest value of $\chi^2$ that would be a rare event (that is, that falls in the upper 5% of the distribution).

Using the Chi-Square Table and Your Calculator

To find the critical values of $\chi^2$ that cut off the upper 5% of the distribution (or another percentage), you can use the $\chi^2$ table (Table C) in the appendix on page 827. A partial table is shown in Display 10.6. In this table, $df$ (degrees of freedom), is equal to the number of categories minus 1.

The degrees of freedom concept is the same as it was in measuring variation from the mean. With $n$ data values, the deviations from the mean sum to 0, so only $n - 1$ of the deviations are free to vary. Similarly, with $k$ categories the deviations between observed and expected frequencies sum to 0, so only $k - 1$ of the deviations are free to vary.

<table>
<thead>
<tr>
<th>$df$</th>
<th>0.10</th>
<th>0.05</th>
<th>...</th>
<th>0.01</th>
<th>0.005</th>
<th>...</th>
<th>0.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.71</td>
<td>3.84</td>
<td>...</td>
<td>6.63</td>
<td>7.88</td>
<td>...</td>
<td>10.83</td>
</tr>
<tr>
<td>2</td>
<td>4.61</td>
<td>5.99</td>
<td>...</td>
<td>9.21</td>
<td>10.60</td>
<td>...</td>
<td>13.82</td>
</tr>
<tr>
<td>3</td>
<td>6.25</td>
<td>7.81</td>
<td>...</td>
<td>11.34</td>
<td>12.84</td>
<td>...</td>
<td>16.27</td>
</tr>
<tr>
<td>4</td>
<td>7.78</td>
<td>9.49</td>
<td>...</td>
<td>13.23</td>
<td>14.86</td>
<td>...</td>
<td>18.47</td>
</tr>
<tr>
<td>5</td>
<td>9.24</td>
<td>11.07</td>
<td>...</td>
<td>15.09</td>
<td>16.75</td>
<td>...</td>
<td>20.51</td>
</tr>
<tr>
<td>6</td>
<td>10.64</td>
<td>12.53</td>
<td>...</td>
<td>16.81</td>
<td>18.55</td>
<td>...</td>
<td>22.46</td>
</tr>
<tr>
<td>7</td>
<td>12.02</td>
<td>14.07</td>
<td>...</td>
<td>18.48</td>
<td>20.28</td>
<td>...</td>
<td>24.32</td>
</tr>
<tr>
<td>8</td>
<td>13.36</td>
<td>15.51</td>
<td>...</td>
<td>20.09</td>
<td>21.95</td>
<td>...</td>
<td>26.12</td>
</tr>
<tr>
<td>9</td>
<td>14.68</td>
<td>16.92</td>
<td>...</td>
<td>21.67</td>
<td>23.59</td>
<td>...</td>
<td>27.83</td>
</tr>
<tr>
<td>10</td>
<td>15.99</td>
<td>18.31</td>
<td>...</td>
<td>23.21</td>
<td>25.19</td>
<td>...</td>
<td>29.59</td>
</tr>
<tr>
<td>11</td>
<td>17.29</td>
<td>19.68</td>
<td>...</td>
<td>24.72</td>
<td>26.76</td>
<td>...</td>
<td>31.26</td>
</tr>
<tr>
<td>12</td>
<td>18.55</td>
<td>21.03</td>
<td>...</td>
<td>26.22</td>
<td>28.30</td>
<td>...</td>
<td>32.91</td>
</tr>
</tbody>
</table>

Display 10.6 A partial table of critical values of $\chi^2$ from Table C.
**Example: Critical Values and Approximate P-Values**

In the fair die example, the observed value of the $\chi^2$ statistic was 2.6. Would you reject the null hypothesis that the die is fair at the 5% level of significance? What is the approximate P-value?

**Solution**

To find the value of $\chi^2$ that cuts off the upper 5% of the distribution when $\chi^2$ is computed for rolling a fair six-sided die, go to the row $df = 6 - 1$, or 5, in Display 10.6. Then go over to the column headed .05. The critical value, $\chi^2*$, is 11.07. The computed value of $\chi^2$ from the fair die example is only 2.6, so you have no evidence at the $\alpha = 0.05$ level that the die is unfair. As illustrated in Display 10.7, it is quite likely that you would get a value of 2.6 or greater with a fair die.

Looking at the row of values for 5 degrees of freedom in Display 10.6, you see that the smallest value is 9.24, which cuts off an upper tail area of 0.10. The observed value, 2.6, is well to the left of 9.24 under the $\chi^2$ distribution (see Display 10.7). Therefore, all you can say about the P-value is that it is larger than 0.10. (Table C on page 827 tells you that it is also larger than 0.25.) Statistical software or a calculator can give you an exact P-value. [To learn how to find the P-value with your calculator, see **Calculator Note 10D**.] Here, the P-value is approximately 0.76.

**The Chi-Square Goodness-of-Fit Test**

So far in this section, you have used the fair die example to develop the ideas of a chi-square goodness-of-fit test. However, you can apply this test in important, real-life situations in which you wish to assess how well a given probability model fits your data. To perform a chi-square goodness-of-fit test, you should go through the same steps as for any test of significance.
1. **Name the test and check conditions.**
   - Each outcome in your population falls into exactly one of a fixed number of categories.
   - You have a model that gives the hypothesized proportion of outcomes in the population that fall into each category.
   - You have a random sample from your population.
   - The expected frequency in each category is 5 or greater.

2. **State the hypotheses.**
   - \( H_0 \): The proportions in the population are equal to the proportions in your model.
   - \( H_a \): At least one proportion in the population is not equal to the corresponding proportion in your model.

3. **Compute the value of the test statistic, approximate the \( P \)-value, and draw a sketch.** The test statistic is
   \[
   \chi^2 = \sum \frac{(O - E)^2}{E}
   \]
   where \( O \) is the observed frequency in each category and \( E \) is the corresponding expected frequency.

   The \( P \)-value is the probability of getting a value of \( \chi^2 \) as extreme as or even more extreme than the one in the sample, assuming the null hypothesis is true. Get a \( P \)-value from your calculator, or approximate the \( P \)-value by comparing the value of \( \chi^2 \) to the appropriate critical value of \( \chi^2 \) in Table C, where \( df = \text{number of categories} - 1 \).

4. **Write your conclusion linked to your computations and in the context of the problem.** If the \( P \)-value is smaller than \( \alpha \) (or, equivalently, if \( \chi^2 \) is larger than the critical value for the given \( \alpha \)), then reject the null hypothesis. If not, then you don't have statistically significant evidence that the null hypothesis is false and so you do not reject it. (Remember that you don't say you “accept” the null hypothesis. This is because you don't know that it is true; you simply don't have any evidence that the hypothesized proportions are exactly the right ones, only evidence that your data are consistent with those proportions. Your data will be consistent with other proportions, too.)
Chapter 10 Chi-Square Tests

Example: Milk Chocolate Versus Peanut Butter

Milk chocolate M&M’s are 13% red, 14% yellow, 16% green, 20% orange, 13% brown, and 24% blue. A random sample of 200 peanut butter M&M’s yielded the distribution of colors shown in Display 10.8. Do you have evidence that the distribution of peanut butter M&M’s is different from the distribution of milk chocolate M&M’s?

<table>
<thead>
<tr>
<th>Color</th>
<th>Observed Number of M&amp;M’s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red</td>
<td>25</td>
</tr>
<tr>
<td>Yellow</td>
<td>37</td>
</tr>
<tr>
<td>Green</td>
<td>45</td>
</tr>
<tr>
<td>Orange</td>
<td>34</td>
</tr>
<tr>
<td>Brown</td>
<td>19</td>
</tr>
<tr>
<td>Blue</td>
<td>40</td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
</tr>
</tbody>
</table>

Display 10.8 Distribution of colors in a sample of peanut butter M&M’s.

Solution

If the distribution of colors for peanut butter M&M’s is the same as that for milk chocolate M&M’s, you would expect the numbers of each color to be as shown in Display 10.9. Note that in a chi-square test it is not necessary that all of the expected frequencies be the same.

<table>
<thead>
<tr>
<th>Color</th>
<th>Percentage in Milk Chocolate M&amp;M’s</th>
<th>Expected Number of M&amp;M’s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red</td>
<td>13%</td>
<td>0.13(200) = 26</td>
</tr>
<tr>
<td>Yellow</td>
<td>14%</td>
<td>0.14(200) = 28</td>
</tr>
<tr>
<td>Green</td>
<td>16%</td>
<td>0.16(200) = 32</td>
</tr>
<tr>
<td>Orange</td>
<td>20%</td>
<td>0.20(200) = 40</td>
</tr>
<tr>
<td>Brown</td>
<td>13%</td>
<td>0.13(200) = 26</td>
</tr>
<tr>
<td>Blue</td>
<td>24%</td>
<td>0.24(200) = 48</td>
</tr>
<tr>
<td>Total</td>
<td>100%</td>
<td>200</td>
</tr>
</tbody>
</table>

Display 10.9 Expected number of M&M’s of different colors.

[Source: us.mms.com, March 2006.]

Check conditions.

The conditions for a chi-square goodness-of-fit test are met in this situation. You have a random sample of 200 peanut butter M&M’s. Each M&M was only one color. You can compute the expected number of each color because you know the distribution of colors in milk chocolate M&M’s. All of the expected counts are at least 5.

State the hypotheses.

The null hypothesis is

H₀: The distribution of colors in peanut butter M&M’s is the same as the distribution of colors in milk chocolate M&M’s; that is, there are 13% red, 14% yellow, 16% green, 20% orange, 13% brown, and 24% blue.
The alternative hypothesis is

\[ H_a: \text{The distribution of colors in peanut butter M&M's is not the same as the distribution of colors in milk chocolate M&M's; that is, at least one proportion is different.} \]

The test statistic is

\[
\chi^2 = \sum \frac{(O - E)^2}{E}
\]

\[
= \frac{(25 - 26)^2}{26} + \frac{(37 - 28)^2}{28} + \frac{(45 - 32)^2}{32} + \frac{(34 - 40)^2}{40} + \frac{(19 - 26)^2}{26} + \frac{(40 - 48)^2}{48}
\]

\[ \approx 12.33 \]

The value of \( \chi^2 \) from the sample, 12.33, is quite far out in the tail of the chi-square distribution with \( 6 - 1 \), or 5, degrees of freedom. In fact, it is in the upper 0.031 of the tail.

Reject the null hypothesis. You cannot attribute the difference in the expected and observed frequencies to variation in sampling alone. A value of \( \chi^2 \) this large is very unlikely to occur in random samples of this size if peanut butter M&M’s have the same distribution of colors as milk chocolate M&M’s. Conclude that the distribution of colors in peanut butter M&M’s is different from that in milk chocolate M&M’s. Display 10.10 shows a Fathom printout for this test.
The Chi-Square Goodness-of-Fit Test

D9. In the example of milk chocolate versus peanut butter M&M’s, does the final value of the \( \chi^2 \) statistic tell you where the lack of fit occurs (i.e., which colors might be more prevalent or less prevalent in the peanut butter candies)? If not, where can you look for such information?

D10. Which of these statements describe properties of the chi-square goodness-of-fit test?

A. If you switch the order of the categories, the value of the \( \chi^2 \) statistic does not change.
B. The “observed” frequencies are always whole numbers.
C. The “expected” frequencies are always whole numbers.
D. The number of degrees of freedom is 1 less than the sample size.
E. A high value of \( \chi^2 \) indicates a high level of agreement between the observed frequencies and the expected frequencies.

D11. Why is a \( \chi^2 \) test typically one-sided toward the large values?

Why Must Each Expected Frequency Be 5 or More?

The theoretical \( \chi^2 \) distribution used in the table is a continuous distribution. For example, Display 10.11 shows a (continuous) \( \chi^2 \) distribution for ten categories (\( df = 9 \)) and another for six categories (\( df = 5 \)).

![chi-square distributions](image)

Display 10.11 Chi-square distributions for \( df = 5 \) and \( df = 9 \).

The distribution of \( \chi^2 \) computed from repeated sampling is discrete, however, because only a limited number of distinct values of \( \chi^2 \) can be calculated for a given number of categories and a given sample size. Like the normal approximation to the binomial distribution, the \( \chi^2 \) distribution is a continuous distribution that can be used to approximate a discrete distribution. The larger the expected frequencies, the closer the distribution of possible values of \( \chi^2 \) is to a continuous distribution. In order to have a reasonable approximation, the expected frequency in each category should be 5 or greater. (This is a conservative rule, but it works well in most cases.)
When the Chi-Square Test Is Equivalent to the z-Test

The chi-square test can be thought of as an extension of the z-test to more than two categories. Suppose you were checking a coin, rather than a die, to see if it was fair. If you spin the coin \( n \) times, you would expect to see \( \frac{n}{2} \) heads under the hypothesis that the spinning coin is fair. Letting \( x_1 \) denote the number of heads in the sample of \( n \) spins and \( x_2 \) the number of tails, the table for constructing a \( \chi^2 \) statistic for the data would look like Display 10.12.

<table>
<thead>
<tr>
<th>Category</th>
<th>Observed Frequency, ( O )</th>
<th>Expected Frequency, ( E )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heads</td>
<td>( x_1 ) ( \frac{n}{2} )</td>
<td></td>
</tr>
<tr>
<td>Tails</td>
<td>( x_2 ) ( \frac{n}{2} )</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>( n ) ( n )</td>
<td></td>
</tr>
</tbody>
</table>

Display 10.12 Table for computing \( \chi^2 \) when testing the fairness of a coin.

The test statistic for testing the null hypothesis that spinning the coin is fair becomes

\[
\chi^2 = \frac{(x_1 - \frac{n}{2})^2}{\frac{n}{2}} + \frac{(x_2 - \frac{n}{2})^2}{\frac{n}{2}}
\]

The sum of the number of heads and tails must be \( n \), so \( x_2 = n - x_1 \) and you can write the test statistic as

\[
\chi^2 = \frac{(x_1 - \frac{n}{2})^2}{\frac{n}{2}} + \frac{\left(n - x_1\right)^2 - \frac{n}{2}}{\frac{n}{2}} = 2 \frac{(x_1 - \frac{n}{2})^2}{\frac{n}{2}} = \frac{(x_1 - \frac{n}{2})^2}{n \left(\frac{1}{2}\right)}
\]

But you already know another way of testing the hypothesis that spinning a coin is fair. For large \( n \), the test statistic for this hypothesis could be the familiar z-statistic from Chapter 8, given by

\[
z = \frac{\hat{p} - p}{\sqrt{pq/n}} = \frac{x_1/n - \frac{1}{2}}{\sqrt{\left(\frac{1}{2}\right)o{n}\left(\frac{1}{2}\right)}} \approx \frac{x_1 - \frac{n}{2}}{\sqrt{\left(\frac{1}{2}\right)n}}
\]

You can now see that \( \chi^2 = z^2 \). The square of a z-statistic has a \( \chi^2 \) distribution with 1 degree of freedom. This equality holds in general, even if \( p \) isn’t \( \frac{1}{2} \).

In summary, if there are only two types of outcomes (success and failure), you are back in the binomial situation and the z-test from Chapter 8 is equivalent to the chi-square test of this chapter. This implies, among other things, that the assumptions for the chi-square test are the same as those for the z-test: The sample must be random and be large enough so that the sample proportion has an approximately normal sampling distribution.
When the Chi-Square Test Is Equivalent to the z-Test

D12. Discuss the similarities and differences between a z-test using proportions and a chi-square test using frequencies.

D13. It is hypothesized that among the students at your school who will be seniors next year, 20% will buy their lunch on campus, 30% will carry their own lunch, and 50% will eat off campus. A random sample of these students will be surveyed to check the claim that only half will eat off campus. How would you construct this test? Defend your choice.

Summary 10.1: Testing a Probability Model: The Chi-Square Goodness-of-Fit Test

The chi-square goodness-of-fit test is used when

- each outcome in your population falls into exactly one of a fixed number of categories
- you have a random sample from your population
- you want to know if it’s plausible that the proportion of outcomes in the population that fall into each category is equal to the corresponding proportion in some hypothesized model.

To perform a chi-square goodness-of-fit test, you go through the same steps as for any test of significance (see page 681). The expected frequencies are calculated from the probabilities specified in the null hypothesis, and

\[ df = \text{number of categories} - 1. \]

If the results from your sample are very different from the results expected from the model, then \( \chi^2 \) will be large and you will reject the hypothesis that the model is the correct one for your population. If the observed frequencies from your sample are close to the expected frequencies, then \( \chi^2 \) will be relatively small and you have no evidence that the model is incorrect.

In the exercises, you will see examples of how the chi-square goodness-of-fit test is used in the real world.

Practice

A Test Statistic

P1. Suppose you want to test whether a tetrahedral (four-sided) die is fair. You roll it 50 times and observe a one 14 times, a two 17 times, a three 9 times, and a four 10 times.
   a. How many do you “expect” in each category?
   b. Compute the value of \( \chi^2 \).

P2. If each observed frequency equals the expected frequency, what is the value of \( \chi^2 \)?

The Distribution of Chi-Square

P3. Refer to Display 10.5 on page 678. Suppose you roll a 12-sided die 300 times to see if it is fair. You compute \( \chi^2 \) and get 21.3. Use the appropriate histogram to approximate a P-value for this test. What is your conclusion if you are using \( \alpha = 0.05 \)?

P4. Refer to Display 10.5 on page 678. Suppose you roll an eight-sided die 100 times to see if it is fair, and get \( \chi^2 = 21.3 \). Use the appropriate histogram to approximate a P-value for this test. What is your conclusion if \( \alpha = 0.05 \)?
Using the Chi-Square Table and Your Calculator

P5. Give \( df \) for each situation, and then use Table C on page 827 to determine if the result is statistically significant (\( \alpha = 0.05 \)).
   a. You roll a tetrahedral die 100 times and calculate a \( \chi^2 \) value of 8.24.
   b. You roll a 20-sided die 500 times and calculate a \( \chi^2 \) value of 8.24.

P6. Learn to use your calculator to find \( P \)-values, and then find the \( P \)-value for the case of
   a. rolling an 8-sided die, and \( \chi^2 = 2.6 \)
   b. rolling a 12-sided die, and \( \chi^2 = 21.3 \)

The Chi-Square Goodness-of-Fit Test

P7. For 1000 shoppers donating blood at a local mall, the frequencies of blood types were as shown in Display 10.13.

<table>
<thead>
<tr>
<th>Blood Type</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>O</td>
<td>465</td>
</tr>
<tr>
<td>A</td>
<td>394</td>
</tr>
<tr>
<td>B</td>
<td>96</td>
</tr>
<tr>
<td>AB</td>
<td>45</td>
</tr>
<tr>
<td>Total</td>
<td>1000</td>
</tr>
</tbody>
</table>

Display 10.13 Frequency of blood types of 1000 blood donors at a mall.

Theory says these blood types should be in the ratio 9:8:2:1. For example, the fraction of people having type O blood should be \( \frac{9}{20} \). Do you have statistically significant evidence that the data do not support this model? Be sure to include all four steps of a test of significance. [Source: P. V. Rao, Statistical Methods in the Life Sciences (Pacific Grove, Calif.: Duxbury, 1998), p. 221.]

P8. Many people think that the full moon prompts strange behavior and occurrences. One commonly held belief is that epileptic seizures are more frequent during a full moon. To check this claim, a researcher observed that, among 470 epileptic seizures monitored at Tampa General Hospital, 103 occurred during the new moon, 121 during the first quarter, 94 during the full moon, and 152 during the last quarter. [Source: Los Angeles Times, June 7, 2004, p. F2.]

   a. Are these data consistent with a model of equal likelihood of seizures during the four periods? If not, where are the largest deviations from the model?
   b. What else would you like to know about the time frames in which these data were collected?

P9. Display 10.14, repeated from Chapter 7, gives the distribution of the number of children per family in the United States.

<table>
<thead>
<tr>
<th>Number of Children in Family</th>
<th>Proportion of Families</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.524</td>
</tr>
<tr>
<td>1</td>
<td>0.201</td>
</tr>
<tr>
<td>2</td>
<td>0.179</td>
</tr>
<tr>
<td>3</td>
<td>0.070</td>
</tr>
<tr>
<td>4 (or more)</td>
<td>0.026</td>
</tr>
</tbody>
</table>


Suppose you want to determine whether the distribution of the number of children in the families of statistics students is consistent with the distribution for families nationwide. Record the number of children in the family of each student in your statistics class. (If you are unable to do this, use the table in Display 10.15, on the next page.) Assume that your class can be considered a random sample of all statistics students.

   a. Do these data satisfy the conditions for a chi-square goodness-of-fit test?
   b. When expected frequencies aren’t large enough, investigators sometimes collapse
several categories into one category. Do that with these data and finish the test.

<table>
<thead>
<tr>
<th>Number of Children in Family</th>
<th>Number of Families</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>4 (or more)</td>
<td>2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>40</strong></td>
</tr>
</tbody>
</table>

**Display 10.15** Observed frequencies of the number of children for 40 families.

When the Chi-Square Test Is Equivalent to the z-Test

P10. Answer the question in the introduction to this chapter: You spin a coin 100 times and get tails 64 times. Do you have evidence that spinning a coin is unfair?
   a. Use inference for a proportion, as in Chapter 8.
   b. Use a chi-square goodness-of-fit test.
   c. Compare the two results in parts a and b.

P11. To demonstrate the differences between flipping and spinning, Professors A. Gelman and D. Nolan modified a checker by adding putty to the crown side, calling it heads. In 100 flips of the checker, they observed 54 heads. In 100 spins of the checker, they observed 23 heads. The results for the flips are consistent with the model that the probability of heads is 0.5, even with the putty. Are the results for the spins of checkers with putty consistent with the model that the probability of heads is 0.5? [Source: stat-www.berkeley.edu]
   a. Answer the question by using inference for a proportion, as in Chapter 8.
   b. Answer the question by using a chi-square goodness-of-fit test.
   c. Compare the two results in parts a and b.

**Exercises**

E1. In 1882, R. Wolf rolled a die 20,000 times. The results are recorded in Display 10.16. Is this evidence that the die was unfair, or is it approximately what you would expect from a fair die?

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3407</td>
</tr>
<tr>
<td>2</td>
<td>3631</td>
</tr>
<tr>
<td>3</td>
<td>3176</td>
</tr>
<tr>
<td>4</td>
<td>2916</td>
</tr>
<tr>
<td>5</td>
<td>3448</td>
</tr>
<tr>
<td>6</td>
<td>3422</td>
</tr>
</tbody>
</table>

**Display 10.16** Results from 20,000 rolls of a die. [Source: D. J. Hand et al., A Handbook of Small Data Sets (London: Chapman & Hall, 1994), p. 29.]

E2. Gelman and Nolan had students roll a die that had the corners on the 1 side slightly rounded. The results of 120 rolls are shown in Display 10.17.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>41</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>3</td>
<td>28</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>22</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>

**Display 10.17** Frequencies of outcomes of 120 rolls of a shaved die. [Source: stat-www.berkeley.edu]

Clearly, there are more 1’s than you would expect and fewer 6’s, the outcome on the side opposite the 1. Is there evidence that the shaved die was unfair with regard to the other four outcomes?

E3. A study attempted to find a relationship between people’s birthdays and dates of admission for treatment of alcoholism. In a sample of 200 admissions, 11 were within 7 days of the person’s birthday; 24 were between 8 and 30 days, inclusive; 69 were
between 31 and 90 days, inclusive; and 96 were more than 90 days from the person’s birthday. Find appropriate expected numbers of admissions, and perform a chi-square goodness-of-fit test to test the hypothesis that admission is unrelated to a person’s birthday. [Source: Jay Devore and Rossy Peck, Statistics: The Exploration and Analysis of Data, 3d ed. (Belmont, Calif.: Duxbury Press, 1997), pp. 565–66. Original source: Psychological Reports, 1992, pp. 944–46.]

E4. Display 10.18 gives the number of births for each month in a hospital in Switzerland. Only the 700 women who were having their first baby were included. Is there evidence that first births are not spread evenly throughout the year?

<table>
<thead>
<tr>
<th>Month</th>
<th>Number of Births</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>66</td>
</tr>
<tr>
<td>February</td>
<td>63</td>
</tr>
<tr>
<td>March</td>
<td>64</td>
</tr>
<tr>
<td>April</td>
<td>48</td>
</tr>
<tr>
<td>May</td>
<td>64</td>
</tr>
<tr>
<td>June</td>
<td>74</td>
</tr>
<tr>
<td>July</td>
<td>70</td>
</tr>
<tr>
<td>August</td>
<td>59</td>
</tr>
<tr>
<td>September</td>
<td>54</td>
</tr>
<tr>
<td>October</td>
<td>51</td>
</tr>
<tr>
<td>November</td>
<td>45</td>
</tr>
<tr>
<td>December</td>
<td>42</td>
</tr>
</tbody>
</table>


E5. Gregor Mendel (Austrian, 1822–1884) performed experiments to try to validate the predictions of his genetic theory. Mendel’s experimental results match his predictions quite closely. In fact, statisticians think his results match too closely. In one experiment, Mendel predicted that he would get a 9:3:3:1 ratio between smooth yellow peas, wrinkled yellow peas, smooth green peas, and wrinkled green peas. His experiment resulted in 315 smooth yellow peas, 101 wrinkled yellow peas, 108 smooth green peas, and 32 wrinkled green peas. Are the observed counts suspiciously close to the counts predicted by Mendel’s theory?

E6. The characteristics of the 529 offspring of the plants discussed in E5 are given in Display 10.19.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td>38</td>
</tr>
<tr>
<td>Ab</td>
<td>35</td>
</tr>
<tr>
<td>aB</td>
<td>28</td>
</tr>
<tr>
<td>ab</td>
<td>30</td>
</tr>
<tr>
<td>ABb</td>
<td>65</td>
</tr>
<tr>
<td>aBb</td>
<td>68</td>
</tr>
<tr>
<td>AaB</td>
<td>60</td>
</tr>
<tr>
<td>Aab</td>
<td>67</td>
</tr>
<tr>
<td>AaBb</td>
<td>138</td>
</tr>
</tbody>
</table>

A = round, a = wrinkled, B = yellow, and b = green

Display 10.19 Frequencies of the genotypes of 529 pea plants. [Source: www.mendelweb.org.]

According to Mendel’s theory, these genotypes, starting from the top, should be in the ratio 1:1:1:2:2:2:2:4. Do the data support the theory? Do they support it better than you would expect?

E7. One of the most famous U.S. elections of all time (prior to 2000) was the Truman–Dewey election of 1948. It became famous because Dewey was predicted to win by almost everyone and by almost every poll. Display 10.20 gives the final predictions (in percentage of votes) of the three most popular national polls of the day, along with the actual result.

<table>
<thead>
<tr>
<th>Candidate</th>
<th>Crossley</th>
<th>Gallup</th>
<th>Roper</th>
<th>Actual Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truman</td>
<td>45</td>
<td>44</td>
<td>38</td>
<td>50</td>
</tr>
<tr>
<td>Dewey</td>
<td>50</td>
<td>50</td>
<td>53</td>
<td>45</td>
</tr>
<tr>
<td>Thurmond</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Wallace</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

Suppose each poll sampled about 3000 voters. Do any of the three poll results demonstrate a reasonably good fit to the actual results? Which of the three fits best?

E8. The 1948 presidential election polls of Crossley, Gallup, and Roper (see E7) were based on a method called quota sampling, in which quotas such as so many males, so many females, so many retired, so many working, so many unemployed, and so on were to be filled by field-workers for the polling company. The state of Washington, however, tried out a fairly new method called “probability sampling,” which made use of the random sampling ideas used today. Display 10.21 shows how that poll, based on about 1000 voters, came out, in terms of percentages.

<table>
<thead>
<tr>
<th>Probability Sample</th>
<th>Actual Vote</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dewey</td>
<td>46.0</td>
</tr>
<tr>
<td>Truman</td>
<td>50.5</td>
</tr>
<tr>
<td>Wallace</td>
<td>2.9</td>
</tr>
<tr>
<td>Other</td>
<td>0.6</td>
</tr>
</tbody>
</table>


Can you reject the hypothesis that this sample result is a good fit to the “truth”?  

E9. Does the number of people involved in traffic accidents increase after a Super Bowl telecast? To answer this question, investigators looked at the number of fatalities on public roadways in the United States for the first 27 Super Bowl Sundays. They compared the number of fatalities during the 4 hours after the telecast with the number during the same time period on the Sundays the week before and the week after the Super Bowl (54 control Sundays). If the telecast had no effect, there should be roughly twice the number of fatalities on the 54 control Sundays as on the 27 Super Bowl Sundays. [Source: D. A. Redelmeier and C. L. Stewart, “Do Fatal Crashes Increase Following a Super Bowl Telecast?” Chance 18, no. 1 (2003): 19–24.]

a. There were 662 fatalities in the 4 hours following Super Bowl telecasts and 936 in the corresponding hours on control Sundays. Are these data consistent with the model that the telecast has no effect? What can you conclude?

b. Within the home state of the winning Super Bowl team, the number involved in traffic accidents (fatalities plus survivors) totaled 141 for Super Bowl Sundays and 265 for control Sundays. Do these data fit the model given? What can you conclude?

E10. It is sometimes said that older people are overrepresented on juries. Display 10.22 gives information about people on grand juries in Alameda County, California. Does it appear that these grand jurors were selected at random from the adult population of Alameda County?

<table>
<thead>
<tr>
<th>Age</th>
<th>Countywide Percentage</th>
<th>Number of Grand Jurors</th>
</tr>
</thead>
<tbody>
<tr>
<td>21–40</td>
<td>42</td>
<td>5</td>
</tr>
<tr>
<td>41–50</td>
<td>23</td>
<td>9</td>
</tr>
<tr>
<td>51–60</td>
<td>16</td>
<td>19</td>
</tr>
<tr>
<td>61 or older</td>
<td>19</td>
<td>33</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>66</td>
</tr>
</tbody>
</table>


E11. Write a research hypothesis about data that can be analyzed using a chi-square goodness-of-fit test. For example, “People are less likely to be born on some days of the week than on others.” Design and carry out a survey about your hypothesis using a random sample of a specified population.

E12. A sign on a barrel of nuts in a supermarket says that it contains 30% cashews, 30% hazelnuts, and 40% peanuts by weight. You mix up the nuts and scoop out 20 lb. When
you weigh the nuts, you find that you have 6 lb of cashews, 5 lb of hazelnuts, and 9 lb of peanuts.

a. Do you have evidence to doubt the supermarket’s claim?
b. If you have worked this problem using the chi-square test, convert everything to ounces and recalculate.
c. What is your conclusion now?
d. Do you see any problem with using a chi-square test on data like these?

E13. One of the basic principles you have learned in this textbook is that, all else being equal, it is better to have a larger sample than a smaller one.

a. Explain why a larger sample is better.
b. As long as each expected frequency is at least 5, is there any advantage to having a larger sample in a chi-square goodness-of-fit test?

e14. You can use statistical software to simulate the distribution of $\chi^2$ values for any given number of degrees of freedom. Display 10.23 shows values of a $\chi^2$ statistic for 2 degrees of freedom. Each dot represents 9 points.

Display 10.23  Simulated $\chi^2$ distribution for $df = 2$.

Use software to construct a distribution of $\chi^2$ values for various numbers of degrees of freedom. Describe how these distributions change as the number of degrees of freedom increases.

E15. The World Almanac and Book of Facts lists 104 “Major Rivers in North America” with their lengths in miles. These data come from the U.S. Geological Survey. For example, the length of the Hudson River is given as 306 mi, the Columbia as 1243 mi, and the Missouri as 2315 mi. The final digits of the lengths of these three rivers are 6, 3, and 5, respectively.

The Missouri River

The distribution of the final digits is shown in Display 10.24. Test the hypothesis that the digits are equally likely. If they do not appear to be equally likely, what possible explanation can you offer?

<table>
<thead>
<tr>
<th>Final Digit</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>44</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
</tr>
</tbody>
</table>

Display 10.24  Frequencies of the final digits of the lengths of the 104 major rivers of North America.

E16. If you look at the leading digits in a table of data, you might expect each digit to occur about $\frac{1}{9}$ of the time. (There are only nine possibilities because the digit 0 is never a leading digit.) However, smaller digits tend to occur more frequently than larger digits.
Benford’s Law says that the digit $k$ occurs with relative frequency
\[
\log_{10}\left(\frac{k+1}{k}\right)
\]
a. Make a chart that shows the relative frequencies expected for the digits 1 through 9.
b. Prove algebraically that the sum of the relative frequencies is 1.

c. Display 10.25 shows the percentages of first digits seen in a sample of 100 numbers appearing on the first page of various newspapers, collected by Dr. Frank Benford himself. Do they obey his law?

<table>
<thead>
<tr>
<th>Digit</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
</tr>
</tbody>
</table>

Display 10.25  [Source: mathworld.wolfram.com.]

### 10.2 The Chi-Square Test of Homogeneity

A chi-square test of homogeneity tests whether it is reasonable to believe that when several different populations are sorted into the same categories, they will have the same proportion of members in each category. Such populations are called **homogeneous**. Before learning the formal procedure, you will work through some data from a student project.

### Categorical Data with Two Variables

For her project, Justine decided to test the dry strength of three brands of paper towels. Getting a random sample of towels from each brand is pretty much impossible on a student’s budget, but she thought it was reasonable to assume that all towels of the same brand are pretty much identical. So Justine bought one roll of each of the three brands and used the first 25 towels from each. With the help of two friends, she stretched each towel tightly, dropped a golf ball on it from a height of 12 in., and recorded whether the golf ball went through the towel. Justine tested the 75 towels in random order because she realized that it would be impossible to hold her testing procedure completely constant. Display 10.26 shows the results from Justine’s tests.

![Image of Justine testing paper towels]

<table>
<thead>
<tr>
<th>Towel Breaks?</th>
<th>Wipe-Ups</th>
<th>Wipe-Its</th>
<th>Wipe-Outs</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>18</td>
<td>7</td>
<td>5</td>
<td>30</td>
</tr>
<tr>
<td>No</td>
<td>7</td>
<td>18</td>
<td>20</td>
<td>45</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>25</strong></td>
<td><strong>25</strong></td>
<td><strong>25</strong></td>
<td><strong>75</strong></td>
</tr>
</tbody>
</table>

Display 10.26  Two-way table of observed frequencies for Justine’s paper towels data.

The table in Display 10.26 is called a **two-way table** because each outcome is classified in two ways: according to the brand of paper towel (the population it comes from) and whether or not it breaks. Disregarding the labels and totals, it has two rows, three columns, and six cells. The data are called “categorical” because the only information recorded about each response is which category
(or class) it falls into. A cell of the table consists of a count (or frequency). For example, the 20 in the second row and third column is the number of paper towels in Justine’s sample of the 25 Wipe-Outs that did not break.

The column totals, indicating 25 towels of each brand, and the row totals, indicating 30 towels that broke and 45 that did not break, are called the **marginal frequencies**. The summary that \( \frac{30}{75} \) or 0.40, or 40%, of the towels broke is called a **marginal relative frequency**. A total of 18 of the Wipe-Up towels broke; this is called a **joint frequency** (joint between the column variable, *brand*, and the row variable, *break*). Among the Wipe-Ups, \( \frac{18}{25} \) or 0.72, or 72%, of the towels broke; among the Wipe-Outs, only \( \frac{5}{25} \) or 0.20, or 20%, broke. These are called **conditional relative frequencies** for these columns. The conditional relative frequency for the first row tells you that, among the towels that broke, \( \frac{18}{30} \) or 0.60, or 60%, were Wipe-Ups.

Are the results from Justine’s three samples consistent with the null hypothesis that the percentage of towels that break is the same for all three brands?

As in any analysis, you should look first at a graphical display of the data. One possible plot is given in Display 10.27. This plot is called a **stacked** or **segmented bar chart**. For each population, the observed frequencies in each category are stacked on top of each other. From this plot alone, it’s clear that Wipe-Ups are more likely to break than are the other two brands.

![A segmented bar chart stacks the categorical frequencies on top of each other.](image)

**Display 10.27** A stacked bar chart for Justine’s paper towels data.

**Categorical Data with Two Variables**

D14. Suppose the probability that a paper towel breaks is the same for all three brands.

a. Using Justine’s data in Display 10.26, what is your best estimate of the probability that a paper towel will break?

b. Use your answer to part a to construct a two-way table for the expected results from this experiment under the hypothesis that all three brands have the same probability of breaking.
c. What would the stacked bar chart look like if it were true that the probability that a paper towel will break is the same for all three brands and if the results in the sample happened to come out exactly as expected?

**Computing Expected Frequencies**

Brad tested 32 Wipe-Ups, 18 Wipe-Its, and 25 Wipe-Outs, and his results are shown in Display 10.28.

<table>
<thead>
<tr>
<th>Brand</th>
<th>Wipe-Ups</th>
<th>Wipe-Its</th>
<th>Wipe-Outs</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>18</td>
<td>7</td>
<td>5</td>
<td>30</td>
</tr>
<tr>
<td>No</td>
<td>14</td>
<td>11</td>
<td>20</td>
<td>45</td>
</tr>
<tr>
<td>Total</td>
<td>32</td>
<td>18</td>
<td>25</td>
<td>75</td>
</tr>
</tbody>
</table>

**Display 10.28**  Two-way table of observed frequencies for Brad's paper towels data.

If the null hypothesis that the proportion of towels that break is the same for all three brands of paper towels is true, what is the expected frequency in each cell?

The best estimate of the overall proportion that will break is the marginal relative frequency

\[
\frac{\text{total number that break}}{\text{total number of towels tested}} = \frac{\text{row total for ”yes”}}{\text{grand total}} = \frac{30}{75} = 0.4
\]

The expected frequency of the 32 Wipe-Ups that will break is

\[(\text{overall proportion that break}) \cdot (\text{total number of Wipe-Ups}) = (0.4)(32) = 12.8\]

Following this pattern results in a general formula for expected frequencies.

**Formula for the Expected Frequency, E**

\[
E = \frac{\text{row total}}{\text{grand total}} \cdot (\text{column total}) = \frac{\text{(row total)} \cdot (\text{column total})}{\text{grand total}}
\]

Brad's completed table of expected frequencies appears in Display 10.29.

<table>
<thead>
<tr>
<th>Brand</th>
<th>Wipe-Ups</th>
<th>Wipe-Its</th>
<th>Wipe-Outs</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>(\frac{30}{75} \cdot 32 = 12.8)</td>
<td>(\frac{30}{75} \cdot 18 = 7.2)</td>
<td>(\frac{30}{75} \cdot 25 = 10)</td>
<td>30</td>
</tr>
<tr>
<td>No</td>
<td>(\frac{45}{75} \cdot 32 = 19.2)</td>
<td>(\frac{45}{75} \cdot 18 = 10.8)</td>
<td>(\frac{45}{75} \cdot 25 = 15)</td>
<td>45</td>
</tr>
<tr>
<td>Total</td>
<td>32</td>
<td>18</td>
<td>25</td>
<td>75</td>
</tr>
</tbody>
</table>

**Display 10.29**  Expected frequencies for Brad's paper towels data.
After you finish computing a table of expected frequencies, always check that the row and column totals are the same as in your table of observed frequencies.

**DISCUSSION**

**Computing Expected Frequencies**

D15. A chi-square test of homogeneity looks for evidence against the hypothesis that, for each category, the proportion of the population that falls into the category is the same for all populations. Is the evidence against this claim for the three brands of paper towels stronger for Justine’s data in Display 10.26 or for Brad’s data in Display 10.28? Does the evidence against the hypothesis depend more on marginal relative frequencies or on conditional relative frequencies? Explain your reasoning.

**Computing the Chi-Square Statistic**

A chi-square test of homogeneity can be used to test whether the three brands of towels are equally likely to break. This test is similar to a chi-square test for goodness of fit. The value of $\chi^2$ is computed in the same way, using the observed and expected values in the six cells of the table, but in this test the expected frequencies are estimated from the sample data. In the goodness-of-fit test, the probabilities were specified in the null hypothesis.

**Example: Testing Paper Towels**

Compute $\chi^2$ for Justine’s data in Display 10.26 (page 692).

**Solution**

The work is easily organized in a table, as shown in Display 10.30.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Observed Frequency, O</th>
<th>Expected Frequency, $E$</th>
<th>$(O - E)^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wipe-Up breaks</td>
<td>18</td>
<td>$\frac{30 \times 25}{75} = 10$</td>
<td>8</td>
</tr>
<tr>
<td>Wipe-Up doesn’t break</td>
<td>7</td>
<td>$\frac{45 \times 25}{75} = 15$</td>
<td>-8</td>
</tr>
<tr>
<td>Wipe-It breaks</td>
<td>7</td>
<td>$\frac{30 \times 25}{75} = 10$</td>
<td>-3</td>
</tr>
<tr>
<td>Wipe-It doesn’t break</td>
<td>18</td>
<td>$\frac{45 \times 25}{75} = 15$</td>
<td>3</td>
</tr>
<tr>
<td>Wipe-Out breaks</td>
<td>5</td>
<td>$\frac{30 \times 25}{75} = 10$</td>
<td>-5</td>
</tr>
<tr>
<td>Wipe-Out doesn’t break</td>
<td>20</td>
<td>$\frac{45 \times 25}{75} = 15$</td>
<td>5</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>75</strong></td>
<td><strong>75</strong></td>
<td><strong>0</strong></td>
</tr>
</tbody>
</table>

Display 10.30  Calculating $\chi^2$ for Justine’s paper towels data.
As before,

\[ \chi^2 = \sum \frac{(O - E)^2}{E} \]

which is the sum of the last column, so \( \chi^2 \) is approximately 16.34.

Before you can find the \( P \)-value for this test, you need to determine the number of degrees of freedom, \( df \).

### Degrees of Freedom for a Test of Homogeneity

For a chi-square test of homogeneity, the number of degrees of freedom is

\[ df = (r - 1)(c - 1) \]

where \( r \) is the number of rows in the table of observed values and \( c \) is the number of columns (not counting the headings or totals in either case).

#### Example: Paper Towels, Degrees of Freedom

What is the number of degrees of freedom for Justine’s data in Display 10.26 (page 692)? Find and interpret the \( P \)-value.

**Solution**

There are two rows (whether the towel breaks or not) and three columns (the three brands of paper towels). The number of degrees of freedom is

\[ df = (r - 1)(c - 1) = (2 - 1)(3 - 1) = 2 \]

Using the calculator to determine the \( P \)-value for this test, where \( \chi^2 \approx 16.34 \), you get approximately 0.00028. Alternatively, if you look at \( \chi^2 \) values in Table C on page 827, you will find that this value of \( \chi^2 \) is significant at the 0.001 level, as shown in Display 10.31. The \( P \)-value, 0.00028, means that if the null hypothesis that the three brands are equally likely to break is true, then there is almost no chance of getting observed frequencies as far from the expected frequencies as Justine did. Consequently, you would reject the hypothesis that the probability of breaking is the same for all three brands of paper towels.

![Display 10.31](image)  
\( \chi^2 \) distribution with \( df = 2, \alpha = 0.001 \).
**Procedure for a Chi-Square Test of Homogeneity**

The steps in a chi-square test of homogeneity are much the same as the steps in the chi-square goodness-of-fit test.

**Chi-Square Test of Homogeneity**

1. **Name the test and check conditions.**
   - Independent simple random samples of fixed (but not necessarily equal) sizes are taken from two or more large populations (or two or more treatments are randomly assigned to subjects who give a categorical response).
   - Each outcome falls into exactly one of several categories, with the categories being the same in all populations.
   - The expected frequency in each cell is 5 or greater.

2. **State the hypotheses.**
   - $H_0$: The proportion that falls into each category is the same for every population.
   - $H_1$: For at least one category, it is not the case that each population has the same proportion in that category; that is, in some category, the proportion for at least one population is different from that for another population.

3. **Compute the value of the test statistic, approximate the $P$-value, and draw a sketch.** The test statistic is

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

For each cell in a two-way frequency table, $O$ is the observed frequency (count) and $E$ is the expected frequency where:

$$E = \frac{(\text{row total}) \cdot (\text{column total})}{\text{grand total}}$$

The $P$-value is the probability of getting a value of $\chi^2$ as extreme as or more extreme than the one in the sample, assuming the null hypothesis is true. Get the $P$-value from a calculator, or approximate the $P$-value by comparing your value of $\chi^2$ to the appropriate value of $\chi^2$ in Table C on page 827 with

$$df = (r - 1)(c - 1)$$

where $r$ is the number of categories and $c$ is the number of populations.

(continued)
4. Write your conclusion linked to your computations and in the context of the problem. If the P-value is smaller than $\alpha$ (or, equivalently, if $\chi^2$ is larger than the critical value for the given $\alpha$), then reject the null hypothesis. If not, there is no evidence that the null hypothesis is false, so you do not reject it.

Example: Family Values

The Gallup Organization took a poll on family values in different countries. One question asked was “For you personally, do you think it is necessary or not necessary to have a child at some point in your life in order to feel fulfilled?” Display 10.32 shows the results from five countries, for samples of 1000 adults.

<table>
<thead>
<tr>
<th>Country</th>
<th>U.S.</th>
<th>India</th>
<th>Mexico</th>
<th>Canada</th>
<th>Germany</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Necessary</td>
<td>460</td>
<td>930</td>
<td>610</td>
<td>590</td>
<td>490</td>
<td>3080</td>
</tr>
<tr>
<td>Unnecessary</td>
<td>510</td>
<td>60</td>
<td>380</td>
<td>370</td>
<td>450</td>
<td>1770</td>
</tr>
<tr>
<td>Undecided</td>
<td>30</td>
<td>10</td>
<td>10</td>
<td>40</td>
<td>60</td>
<td>150</td>
</tr>
<tr>
<td>Total</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>5000</td>
</tr>
</tbody>
</table>

Display 10.32 Two-way table of observed frequencies and segmented bar chart for the Gallup poll results. [Source: www.gallup.com, May 2002.]

From the plot, you can see that the percentages do not appear to be the same for each country. The difference between India and the United States seems too great. Test the hypothesis that the proportion of adults who would give each answer is the same for each country.
Solution

Under this hypothesis, the expected frequencies are given in Display 10.33.

<table>
<thead>
<tr>
<th>Country</th>
<th>U.S.</th>
<th>India</th>
<th>Mexico</th>
<th>Canada</th>
<th>Germany</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Necessary</td>
<td>616</td>
<td>616</td>
<td>616</td>
<td>616</td>
<td>616</td>
<td>3080</td>
</tr>
<tr>
<td>Unnecessary</td>
<td>354</td>
<td>354</td>
<td>354</td>
<td>354</td>
<td>354</td>
<td>1770</td>
</tr>
<tr>
<td>Undecided</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>150</td>
</tr>
<tr>
<td>Total</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>5000</td>
</tr>
</tbody>
</table>

Display 10.33  Two-way table of expected frequencies for the Gallup poll results.

Check conditions.

The conditions are met in this situation for a chi-square test of homogeneity. There are five large populations, and a random sample of size 1000 was taken from each population. Each answer falls into exactly one of three categories. All the expected frequencies are at least 5.

State the hypotheses.

The null hypothesis is

\[ H_0: \text{If you could ask all adults, the distribution of answers would be the same for each country.} \]

The alternative hypothesis is

\[ H_a: \text{The distribution of answers is not the same in each of the five countries. That is, in at least one country, the proportion of all adults who would give one of the answers is different from the proportion in another country.} \]

Compute the test statistic, approximate the \( P \)-value, and draw a sketch.

The test statistic is

\[
\chi^2 = \frac{(460 - 616)^2}{616} + \frac{(930 - 616)^2}{616} + \cdots + \frac{(40 - 30)^2}{30} + \frac{(60 - 30)^2}{30}
\]

\[= 628.075 \]

Comparing the test statistic to the \( \chi^2 \) distribution with \( df = (3 - 1)(5 - 1) = 8 \), you can see that the value of \( \chi^2 \) from the sample, 628.075, is extremely far out in the tail and is certainly greater than \( \chi^2 = 26.12 \) (\( \alpha = 0.001 \)). The \( P \)-value is approximately 0.

Your calculator will perform a chi-square test of homogeneity and draw a corresponding sketch. [See Calculator Note 10E.]
Display 10.34 shows a Minitab printout of this test.

Chi-Square Test
Expected counts are printed below observed counts

<table>
<thead>
<tr>
<th></th>
<th>U.S.</th>
<th>India</th>
<th>Mexico</th>
<th>Canada</th>
<th>Germany</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>460</td>
<td>930</td>
<td>610</td>
<td>590</td>
<td>490</td>
<td>3080</td>
</tr>
<tr>
<td></td>
<td>616.00</td>
<td>616.00</td>
<td>616.00</td>
<td>616.00</td>
<td>616.00</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>510</td>
<td>60</td>
<td>380</td>
<td>370</td>
<td>450</td>
<td>1770</td>
</tr>
<tr>
<td></td>
<td>354.00</td>
<td>354.00</td>
<td>354.00</td>
<td>354.00</td>
<td>354.00</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>10</td>
<td>10</td>
<td>40</td>
<td>60</td>
<td>150</td>
</tr>
<tr>
<td></td>
<td>30.00</td>
<td>30.00</td>
<td>30.00</td>
<td>30.00</td>
<td>30.00</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>5000</td>
</tr>
</tbody>
</table>

\[\text{ChiSq} = 39.506 + 160.058 + 0.058 + 1.097 + 25.773 + 68.746 + 244.169 + 1.910 + 0.723 + 26.034 + 0.000 + 13.333 + 13.333 + 3.333 + 30.000 = 628.075\]

\[\text{df} = 8, \ p = 0.000\]

Display 10.34  Minitab printout for the family values example.

Reject the null hypothesis. You cannot attribute the differences to the fact that you have only a random sample of adults from each country and not the entire adult population. A value of \(\chi^2\) this large is extremely unlikely to occur in five random samples of this size if the distribution of answers is the same in each country. Conclude that if you questioned all people in these countries, the distributions of answers would differ among some countries.

Example: Premature Infants

Chronic lung disease is one of the primary long-term pulmonary complications among premature infants. This condition generally impairs growth and can result in poor long-term cardiopulmonary function, an increased susceptibility to infection, and increased risk of abnormal neurological development.

A research hypothesis suggested that the use of inhaled nitric oxide would decrease the incidence of chronic lung disease and death in premature infants. To test this hypothesis, a randomized, double-blind, placebo-controlled study of inhaled nitric oxide versus a placebo in premature infants undergoing mechanical ventilation was designed and conducted. A total of 105 infants were randomly assigned to the nitric oxide treatment group, and 102 infants were assigned to the control group. Demographic and baseline clinical characteristics did not differ significantly between the control group and the group given inhaled nitric oxide. The results for the primary goals—reducing death and chronic lung disease—are shown in Display 10.35.
Do these data provide evidence that the inhaled nitric oxide and the placebo result in different distributions of outcomes?

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Nitric Oxide</th>
<th>Placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death</td>
<td>16</td>
<td>23</td>
</tr>
<tr>
<td>Survival with Chronic Lung Disease</td>
<td>35</td>
<td>42</td>
</tr>
<tr>
<td>Survival without Chronic Lung Disease</td>
<td>54</td>
<td>37</td>
</tr>
<tr>
<td>Total</td>
<td>105</td>
<td>102</td>
</tr>
</tbody>
</table>

Display 10.35 Table of observed frequencies for the premature infant study. [Source: Michael D. Schreiber et al., “Inhaled Nitric Oxide in Premature Infants with the Respiratory Distress Syndrome,” New England Journal of Medicine 349, no. 22 (November 27, 2003).]

Solution

Check conditions.

Treatments were randomly assigned to subjects in large enough numbers so that all expected cell frequencies are at least 5 (see Display 10.36). The conditions are met.

State the hypotheses.

The null hypothesis is

\( H_0: \) If both treatments could be assigned to all subjects, the resulting distributions of outcomes would be the same.

The alternative hypothesis is

\( H_a: \) If both treatments could have been assigned to all subjects, the resulting distributions of outcomes would differ. (The chi-square test itself does not tell how they differ.)

Display 10.36 shows the expected frequencies.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Nitric Oxide</th>
<th>Placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death</td>
<td>19.78</td>
<td>19.22</td>
</tr>
<tr>
<td>Survival with Chronic Lung Disease</td>
<td>39.06</td>
<td>37.94</td>
</tr>
<tr>
<td>Survival without Chronic Lung Disease</td>
<td>46.16</td>
<td>44.84</td>
</tr>
<tr>
<td>Total</td>
<td>105</td>
<td>102</td>
</tr>
</tbody>
</table>

Display 10.36 Table of expected frequencies for the premature infant study.

These frequencies result in a \( \chi^2 \) value of

\[
\chi^2 = \frac{(16 - 19.78)^2}{19.78} + \frac{(23 - 19.22)^2}{19.22} + \frac{(35 - 39.06)^2}{39.06} + \frac{(42 - 37.94)^2}{37.94} \\
+ \frac{(54 - 46.16)^2}{46.16} + \frac{(37 - 44.84)^2}{44.84} \\
= 5.026
\]
Comparing this value to the tabled values of the \( \chi^2 \) distribution with 2 degrees of freedom shows that the observed value of the test statistic is a little smaller than the 5% critical value, 5.99. Thus, the \( P \)-value here is a little larger than 0.05 but smaller than 0.10. A calculator gives a \( P \)-value of 0.08.

This is a borderline case. If you strictly interpret this as a fixed 5% level test, the decision is not to reject the null hypothesis and to say you don’t have statistically significant evidence that treating with nitric oxide is an improvement over the placebo. However, the \( P \)-value is fairly small and gives some evidence that something other than random behavior is going on here. As you will see in P20, this point of view is strengthened if you collapse the death and survival with chronic lung disease categories into one category; both are considered undesirable outcomes by the research physicians. The original article reporting this research ends with a positive recommendation for the nitric oxide treatment under certain conditions. See P21 for more information on this study.

## DISCUSSION

### Procedure for a Chi-Square Test of Homogeneity

D16. Discuss the differences between a chi-square test of homogeneity and a chi-square goodness-of-fit test.

D17. Discuss the similarities and differences between a \( z \)-test of the equality of two proportions and the chi-square test of homogeneity.

### Multiple \( z \)-Tests Versus One Chi-Square Test

The decision that the brands of paper towels were not of equal strength in Justine’s project was based on frequencies and a chi-square test. Suppose, however, that the proportion of towels that broke was given for each brand, as in Display 10.37. It appears that a Wipe-Ups paper towel has a much higher probability of breaking than one from the other two brands.

<table>
<thead>
<tr>
<th>Brand</th>
<th>Proportion of Towels That Break</th>
<th>Sample Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wipe-Ups Population 1</td>
<td>0.72</td>
<td>( p_1 )</td>
</tr>
<tr>
<td>Wipe-Its Population 2</td>
<td>0.28</td>
<td>( p_2 )</td>
</tr>
<tr>
<td>Wipe-Outs Population 3</td>
<td>0.20</td>
<td>( p_3 )</td>
</tr>
</tbody>
</table>

**Display 10.37** Table of sample proportions from Justine’s paper towels data.
If there were only two brands of paper towels, you could use a z-test to test whether they are equally likely to break. For Justine’s data, testing the null hypothesis \( p_1 = p_2 \) will lead to the conclusion that these two proportions differ. This two-sample z-test is equivalent to a chi-square test of homogeneity using only the first two columns of data (a test with 1 degree of freedom). As in goodness-of-fit tests, \( \chi^2 = z^2 \). But there are still two more comparisons to be made: \( p_2 \) versus \( p_3 \) and \( p_1 \) versus \( p_3 \).

Why not use multiple z-tests rather than one chi-square test to determine whether the population proportions are homogeneous? This is a deep question, and the answer relies on complex statistical theory. The simple answer is that you have a greater chance of coming to an erroneous conclusion with multiple tests than you do with only one. A guiding statistical principle for hypothesis testing is to never use more tests than you absolutely need. The chi-square test is a clever way of combining many z-tests into one overall test. However, it is appropriate to use z-tests to compare proportions within a table once overall significance has been established.

**Degrees of Freedom: Information About Error**

**Jodain:** I’ve just been letting this “degrees of freedom” thing slide, but in this section they’ve gone too far.

**Ms. C:** What’s the trouble?

**Jodain:** Well, just for a starter, back when we did the t-test, \( df \) was sample size minus 1. I was clueless why we subtracted 1, but never mind that now. Then, in Section 10.1, we had almost the same formula, number of categories minus 1. But why was it number of categories, not sample size?

**Rest of class:** Now we get this hugely different formula, \( df = (r - 1)(c - 1)! \) (grumble, grumble)

**Ms. C:** Come on; it’s a good question. And there’s going to be yet another formula for \( df \) in the next chapter, on regression.

**Jodain:** You keep saying I’m not supposed to just memorize stuff, but . . .

**Ms. C:** Okay, the entire statistical story is way out there, but I’ll try to explain. It has to do with the amount of information in the error term, or \( SE \). Do you remember \( SE \) from the one-sample t-test?

**Jodain:** Sure, I remember. It’s

\[
\frac{s}{\sqrt{n}} \quad \text{where} \quad s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}
\]

**Rest of class:** Figures! Jodain remembers everything!

**Ms. C:** Think about the deviations, \( (x - \bar{x}) \). There are \( n \) of them. Because the sum of the deviations from the mean is 0, if you know all the deviations but one, you can figure out the last one. So you need only \( n - 1 \) deviations to get all the information about the size of a typical deviation from the mean.
Rest of class: It's kind of like if the probability it'll rain tomorrow is 0.4, then you don't learn anything new if they also tell you that the probability it won't rain is 0.6. You could have already figured that out. So you really have only one piece of information.

Ms. C: Well, that's sort of close.

Jodain: Then the reason we didn't have to bother with df in a z-test for a proportion in Section 8.2 is that we knew what the standard error should be because it depended only on \( p_0 \).

Ms. C: Right. You weren't estimating the error term from the data using a sum of squared deviations. So you didn't need to worry about df.

Rest of class: We like z-tests!

Ms. C: What does this have to do with df equaling the number of categories minus 1 for a chi-square goodness-of-fit test?

Jodain: Hmmm. The \( \chi^2 \) statistic itself looks like one big error term. If you are testing whether a six-sided die is fair, there are six categories and six deviations, \( O \) minus \( E \). I suppose if you know all but one of them, you can figure out that one.

Rest of class: Huh?

Ms. C: Jodain's right. The last deviation is determined by the others because the sum of the deviations from the “center” is always equal to 0:

\[
\sum (O - E) = \sum O - \sum E = n - n = 0
\]

Because the last deviation doesn't give you any new information about the size of the deviations from the center,

\[
df = \text{number of categories} - 1
\]

Jodain: But what about the formula in this section, \( df = (r - 1)(c - 1) \)?

Rest of class: (groan) Jodain, why do you do this to us? (mumble, mumble)

Ms. C: Well, how many deviations, \( O \) minus \( E \), are you using in the formula?

Rest of class: We know! The number of rows times the number of columns!

Ms. C: Great answer! Now all you have to do is figure out how many of these give you new information about the size of \( O \) minus \( E \) and how many are redundant.

Jodain: Well, with Justine's paper towels example, there were 2 times 3, or 6, values of \( O \) minus \( E \). They have to sum to 0 in each row and each column. So, for example, if you know the deviations I've put in this table (Display 10.38), you can figure out the rest, meaning that they don't tell you anything new.
Rest of class: We can be missing an entire row and an entire column of deviations!

Ms. C: That means you have only \((r - 1)\) times \((c - 1)\) independent deviations.

Jodain: Here’s my rule: The concept of \(df\) applies when I need to use a sum of squared deviations from a parameter or parameters in my test statistic and I must estimate the parameter or parameters from the data. I count the number of deviations that are free to vary. This is the number of degrees of freedom.

Ms. C: That covers it for everything you’ll see in this class.

Rest of class: We have a different rule—just give us the formula!

**Summary 10.2: The Chi-Square Test of Homogeneity**

Use the chi-square test of homogeneity when

- you have independent samples from two or more populations
- you can classify the response from each member of the sample into exactly one of several categories
- you want to know if it’s plausible that the proportion that falls into each category is the same for each population

This might sound familiar. In Section 8.4, you learned how to test whether the difference between two proportions is statistically significant. This test is the same as a chi-square test of homogeneity when there are two populations and two categories. When there are more than two populations or more than two categories, the chi-square test of homogeneity can test the differences among all the populations in one test.

In a chi-square test of homogeneity, the expected frequencies are calculated from the sample data, and

\[
E = \frac{(\text{row total}) \cdot (\text{column total})}{\text{grand total}}
\]

The number of degrees of freedom for a two-way table with \(r\) rows and \(c\) columns is

\[
df = (r - 1)(c - 1)
\]
**Practice**

**Categorical Data with Two Variables**

P12. Students in a statistics class surveyed random samples of 50 female students and 50 male students, asking each if they preferred a bath or a shower. Display 10.39 gives the results.

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bath</td>
<td>6</td>
<td>21</td>
<td>27</td>
</tr>
<tr>
<td>Shower</td>
<td>44</td>
<td>29</td>
<td>73</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>50</strong></td>
<td><strong>50</strong></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>

Display 10.39 Table of observed preferences for bath versus shower.

a. Construct a segmented bar chart to help you see whether males or females are more likely to prefer a bath.
b. What is your conclusion?

P13. Refer to the data in P12. Give the meaning of each fraction and tell whether each is a marginal relative frequency, a conditional relative frequency, or neither.

a. $\frac{27}{100}$
b. $\frac{21}{27}$
c. $\frac{27}{100}$
d. $\frac{6}{100}$

P14. Refer to P12. Assuming gender doesn't affect the probability of preferring a bath or a shower, construct a table of expected frequencies for 50 males and 50 females.

P15. Display 10.40 gives the number of observations that fall into each of five categories in random samples from three populations.

<table>
<thead>
<tr>
<th></th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
</tr>
<tr>
<td>A</td>
<td>20</td>
</tr>
<tr>
<td>B</td>
<td>21</td>
</tr>
<tr>
<td>C</td>
<td>18</td>
</tr>
<tr>
<td>D</td>
<td>20</td>
</tr>
<tr>
<td>E</td>
<td>23</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>102</strong></td>
</tr>
</tbody>
</table>

Display 10.40 Table of observed frequencies of observations in five categories in samples from three populations.

c. Compute the expected frequency for each cell if the populations are homogeneous.
d. Does it appear that the three populations have the same proportion of members that fall into any given category?

**Constructing the Chi-Square Statistic**

P16. Compute $\chi^2$ for the bath/shower data in P12. What is the number of degrees of freedom for that table? Estimate and interpret the $P$-value.

P17. Compute $\chi^2$ for the data in P15. What is the number of degrees of freedom for the table? Is $\chi^2$ significant at the 5% level?

**Procedures for a Chi-Square Test of Homogeneity**

P18. Verify the table of expected frequencies in Display 10.33 in the family values example on pages 698–699.

P19. Suppose the value of $\chi^2$ for the family values example on pages 698–699 had turned out to be 11.35. Estimate the $P$-value for such a test. Rewrite the conclusion.

P20. In the premature infants example on pages 700–702, collapse the two undesirable outcomes, death and survival with chronic lung disease, into one category and analyze the resulting $2 \times 2$ table. Is there now a statistically significant difference between the two distributions?
P21. Another important aspect of the study on premature infants was the effect of the nitric oxide treatment on intraventricular hemorrhage (bleeding within the heart). Analyze the results reported in Display 10.41 to see if the two distributions differ significantly. If they do, explain the nature of the difference.

Display 10.41  Table of observed frequencies for premature infants.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Nitric Oxide</th>
<th>Placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Severe</td>
<td>13</td>
<td>24</td>
</tr>
<tr>
<td>Not Severe</td>
<td>27</td>
<td>10</td>
</tr>
<tr>
<td>None</td>
<td>65</td>
<td>68</td>
</tr>
<tr>
<td>Total</td>
<td>105</td>
<td>102</td>
</tr>
</tbody>
</table>

Exercises

Each time you are asked to perform a chi-square test, include all four of the steps given on pages 697–698.

E17. A recent Gallup poll asked the same question the Gallup Organization has asked every year for many years: “What do you think is the most important problem facing this country today?” Display 10.42 shows the percentage responses for three major concerns from 2003 to 2006.

<table>
<thead>
<tr>
<th>Most Important Problem</th>
<th>War</th>
<th>Economy</th>
<th>Health Care</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003</td>
<td>19%</td>
<td>20%</td>
<td>9%</td>
<td>52%</td>
</tr>
<tr>
<td>2004</td>
<td>18%</td>
<td>18%</td>
<td>8%</td>
<td>56%</td>
</tr>
<tr>
<td>2005</td>
<td>22%</td>
<td>8%</td>
<td>5%</td>
<td>65%</td>
</tr>
<tr>
<td>2006</td>
<td>23%</td>
<td>10%</td>
<td>6%</td>
<td>61%</td>
</tr>
</tbody>
</table>

Display 10.42  Results from a Gallup poll: responses for three major concerns, 2003–6. [Source: www.gallup.com.]

a. Assume that 1000 people were surveyed each year, and convert the table to one displaying frequencies. Construct and interpret a segmented bar chart of these frequencies.

b. What are the populations in this case?

c. Perform a chi-square test of homogeneity. Use \( \alpha = 0.05 \).

E18. “Overall, how satisfied are you with the quality of education students receive in kindergarten through grade twelve in the U.S. today—would you say you are completely satisfied, somewhat satisfied, somewhat dissatisfied or completely dissatisfied?” This question was posed to a random sample of about 1000 adults in 2004 and another sample of the same size in 2005. The results are shown in Display 10.43. Was there a significant change in the distribution of results from 2004 to 2005? If so, what was the direction of the shift?


a. Display these data in a two-way table and in a segmented bar chart.

Display 10.43  Results of a poll about satisfaction with the quality of education in 2004 and 2005. [Source: poll.gallup.com, 2005.]

<table>
<thead>
<tr>
<th></th>
<th>Completely Satisfied</th>
<th>Somewhat Satisfied</th>
<th>Somewhat Dissatisfied</th>
<th>Completely Dissatisfied</th>
<th>No Opinion</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>90</td>
<td>370</td>
<td>350</td>
<td>160</td>
<td>30</td>
</tr>
<tr>
<td>2004</td>
<td>100</td>
<td>430</td>
<td>320</td>
<td>130</td>
<td>20</td>
</tr>
</tbody>
</table>
b. Compute the value of the test statistic, 
\( z \), for a test of the difference of two 
proportions. Find and interpret the 
\( P \)-value for this test.

c. Compute the value of \( \chi^2 \) for a test of 
homogeneity. Find and interpret the 
\( P \)-value for this test.

d. Compare the two \( P \)-values from parts b 
and c. Compare the values of \( z^2 \) and \( \chi^2 \).

E20. Obesity is associated with many long-term 
illnesses, such as diabetes and hypertension. 
In a Swedish study, a group of subjects 
who had undergone surgery for obesity 
were followed over a 10-year period and 
compared, over that time, to a matched 
sample of conventionally treated control 
subjects. The matching was done based 
on 18 variables, including age, gender, 
weight, smoking habit, and many other 
health-related variables. Among 517 subjects 
who had surgery to reduce obesity, 
7 developed diabetes. Among 539 subjects 
in the control group, 24 developed diabetes. 
[Source: Lars Sjöström et al., “Lifestyle, Diabetes, and 
Cardiovascular Risk Factors 10 Years after Bariatric Surgery,” 
New England Journal of Medicine 351, no. 26 (December 23, 
2004).]

a. Display the data in a two-way table and in 
a segmented bar chart.

b. Compute the value of the test statistic, 
\( z \), for a test of the difference of two 
proportions. Find and interpret the 
\( P \)-value for this test.

c. Compute the value of \( \chi^2 \) for a test of 
homogeneity. Find and interpret the 
\( P \)-value for this test.

d. Compare the two \( P \)-values in parts b and 
c. Compare the values of \( z^2 \) and \( \chi^2 \).

E21. Suppose an experiment is designed to 
determine the effects of a bandage on the 
amount of pain a child perceives in a skinned 
knee. Of the first 60 children who came to 
a school nurse for a skinned knee, 20 were 
selected at random to receive no bandage, 
20 to receive a skin-colored bandage 
(matched to their skin color), and 20 to 
receive a brightly colored bandage. After 
15 minutes, each child was asked, “Is the 
pain gone, almost gone, or still there?” Of 
the 20 children who got no bandage, 0 said 
the pain was gone, 5 said it was almost 
gone, and 15 said it was still there. Of the 
20 children who got a flesh-colored bandage, 
9 said the pain was gone, 9 said it was almost 
gone, and 2 said it was still there. Of the 
20 children who got a brightly colored 
bandage, 15 said the pain was gone, 2 said it 
was almost gone, and 3 said it was still there. Organize these data into a table, display 
them in a plot, and perform a chi-square test 
of homogeneity.

E22. “How does the amount of crime this year 
compare with the amount last year?” This 
question was asked of a random sample 
of 1010 persons in Great Britain and of an 
independent random sample of 1012 persons 
in the United States. In Great Britain, 
43% said more, 25% said less, 21% said the 
same, and the rest had no opinion. In the 
United States, 47% said more, 33% said 
less, 18% said the same, and the rest had no 
opinion. Is there evidence of a significant 
difference in the response pattern between 
the two countries? Organize these data in a 
table, display them in a plot, and perform a 
chi-square test of homogeneity. [Source: poll. 
gallup.com, 2005.]

E23. Gallup Poll Monthly reported on a survey 
built around the question “What is your 
opinion regarding smoking in public 
places?” On the workplace part of the survey, 
respondents were asked to choose which of 
three policies on smoking in the workplace 
they favored. The percentages of the sample 
making each choice for four different years
are shown in Display 10.44. Display these data in a bar chart. Perform a chi-square test of homogeneity to test if the proportion of people who chose each response was the same across the four years. Assume the sample size was approximately 1000 each year.

<table>
<thead>
<tr>
<th>Year</th>
<th>Designated Area</th>
<th>Ban Altogether</th>
<th>No Restrictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>56%</td>
<td>41%</td>
<td>3%</td>
</tr>
<tr>
<td>2003</td>
<td>61%</td>
<td>36%</td>
<td>3%</td>
</tr>
<tr>
<td>2001</td>
<td>58%</td>
<td>38%</td>
<td>4%</td>
</tr>
<tr>
<td>1999</td>
<td>61%</td>
<td>34%</td>
<td>5%</td>
</tr>
</tbody>
</table>

Display 10.44 Results from Gallup polls on smoking in public places. [Source: poll.gallup.com, 2005.]

E24. Gallup polls in 1997 and 2006 asked a randomly selected national sample of approximately 1000 adults, 18 years and older, this question: “Which of the following statements reflects your view of when the effects of global warming will begin to happen? (1) They have already begun to happen, (2) they will start happening within a few years, (3) they will start happening within your lifetime, (4) they will not happen within your lifetime, but they will affect future generations, or (5) they will never happen.” The responses are shown in Display 10.45. Organize the data, display them in a segmented bar chart, write an appropriate null hypothesis, and perform a chi-square test of homogeneity.

<table>
<thead>
<tr>
<th>Response</th>
<th>1997</th>
<th>2006</th>
</tr>
</thead>
<tbody>
<tr>
<td>Already begun</td>
<td>48%</td>
<td>58%</td>
</tr>
<tr>
<td>Within a few years</td>
<td>3%</td>
<td>5%</td>
</tr>
<tr>
<td>In my lifetime</td>
<td>14%</td>
<td>10%</td>
</tr>
<tr>
<td>Future generations</td>
<td>19%</td>
<td>15%</td>
</tr>
<tr>
<td>Never</td>
<td>9%</td>
<td>8%</td>
</tr>
<tr>
<td>No opinion</td>
<td>7%</td>
<td>4%</td>
</tr>
<tr>
<td>Total</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Display 10.45 Results from a Gallup poll on global warming. [Source: www.pollingreport.com, September 2006.]

E25. How does the United States compare to Canada and Great Britain with regard to crime within households? Recent polls regarding victimization within the households in which the respondents lived were conducted in the three countries; the results are shown in Display 10.46. Do the data show significantly differing crime rates for the three countries surveyed? If so, where do the differences occur?

<table>
<thead>
<tr>
<th>Country</th>
<th>Households Victimized by Nonviolent Crime</th>
<th>Households Victimized by Violent Crime</th>
<th>Households Not Victimized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Great Britain</td>
<td>364</td>
<td>81</td>
<td>565</td>
</tr>
<tr>
<td>Canada</td>
<td>331</td>
<td>50</td>
<td>622</td>
</tr>
<tr>
<td>United States</td>
<td>321</td>
<td>51</td>
<td>640</td>
</tr>
<tr>
<td>Total</td>
<td>1010</td>
<td>1003</td>
<td>1012</td>
</tr>
</tbody>
</table>

Display 10.46 Results of polls regarding crime within households. [Source: poll.gallup.com, 2005.]

E26. How does the United States compare to Canada and Great Britain with regard to crime against individuals? Recent polls regarding victimization of those responding to the survey were conducted in the three countries; the results are shown in Display 10.47. Do the data show significantly differing crime victimization rates for the three countries surveyed? If so, where do the largest differences occur?

<table>
<thead>
<tr>
<th>Country</th>
<th>Individuals Victimized by Nonviolent Crime</th>
<th>Individuals Victimized by Violent Crime</th>
<th>Individuals Not Victimized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Great Britain</td>
<td>25%</td>
<td>4%</td>
<td>71%</td>
</tr>
<tr>
<td>Canada</td>
<td>21%</td>
<td>2%</td>
<td>77%</td>
</tr>
<tr>
<td>United States</td>
<td>21%</td>
<td>3%</td>
<td>76%</td>
</tr>
<tr>
<td>Total Individuals</td>
<td>1010</td>
<td>1003</td>
<td>1012</td>
</tr>
</tbody>
</table>

Display 10.47 Results of polls regarding crime against individuals. [Source: poll.gallup.com, 2005.]
E27. Display 10.48 (at the bottom of the page) contains the response percentages from recent surveys that asked this question of randomly selected adults in three countries: “How often, if ever, do you drink alcoholic beverages such as liquor, wine, or beer—every day, a few times a week, about once a week, less than once a week, only on special occasions such as New Year’s and holidays, or never?” Do the three countries differ with respect to drinking patterns? If so, explain where the main differences occur.

E28. It is generally assumed that men are heavier drinkers of alcoholic beverages than are women. Display 10.49 gives the percentages of adults in independent surveys of about 500 men and 500 women from each of three countries who responded yes to the question of whether they had at least several alcoholic drinks per week.

<table>
<thead>
<tr>
<th></th>
<th>Great Britain</th>
<th>Canada</th>
<th>United States</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>48%</td>
<td>37%</td>
<td>26%</td>
</tr>
<tr>
<td>Women</td>
<td>34%</td>
<td>17%</td>
<td>13%</td>
</tr>
</tbody>
</table>

Display 10.49 Percentages of people who drink regularly, out of 500 men and 500 women. [Source: poll.gallup.com, 2006.]

Describe the test you would use to answer each question. Show the test statistic (with all numbers substituted in) that you would compute.

a. Is the proportion of men who drink regularly significantly greater than the proportion of women who drink regularly in the United States?

b. Are there significant differences among the proportions of men who drink regularly across the three countries?

c. Is there a significant difference between the proportions of women who drink regularly in Canada and in the United States?

E29. Gallup conducted a poll of about 750 Internet users in 2003 and a poll with a different sample of about 750 Internet users in 2005. Display 10.50 gives some of the findings. For example, 68% of those interviewed used the Internet for checking news and weather in 2003, and the percentage jumped to 72% in 2005.

<table>
<thead>
<tr>
<th></th>
<th>2003</th>
<th>2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>Checking News and Weather</td>
<td>68%</td>
<td>72%</td>
</tr>
<tr>
<td>Sending and Reading E-Mail</td>
<td>84%</td>
<td>87%</td>
</tr>
<tr>
<td>Shopping</td>
<td>49%</td>
<td>52%</td>
</tr>
</tbody>
</table>

Display 10.50 Results of polls regarding Internet users’ habits, for samples of 750 people in 2003 and 750 different people in 2005. [Source: poll.gallup.com, 2006.]

a. Can this data table be used to conduct a chi-square test of homogeneity on these proportions across the two years? If so, find and interpret the P-value for this test. If not, explain why not.

b. Could you test for equality of proportions of those sending and reading e-mail across the two years? If so, find and interpret the P-value for this test. If not, explain why not.
E30. Two questions were asked in random order of a random sample of about 1000 adults in a recent poll: “At the school your oldest child attends, do you think there is too much emphasis, the right amount, or too little emphasis on sports? On preparing for standardized tests?” The results are shown in Display 10.51. Can these data be used to test for homogeneity of proportions across the two areas by the methods in this section? If so, find and interpret the $P$-value for this test. If not, explain why not.

E31. Look up the definition of *homogeneous*. Explain what two homogeneous populations would be.

E32. Describe how you can tell, for a given set of data, whether to use a chi-square goodness-of-fit test or a chi-square test of homogeneity.

Display 10.51 Results of a poll of parents’ opinions of the amount of emphasis schools place on sports and standardized test preparation, for a sample of 1000 adults. [Source: poll.gallup.com, 2005.]

<table>
<thead>
<tr>
<th></th>
<th>Too Much Emphasis</th>
<th>Right Amount</th>
<th>Too Little Emphasis</th>
<th>Not Applicable</th>
<th>No Opinion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sports</td>
<td>18%</td>
<td>57%</td>
<td>20%</td>
<td>4%</td>
<td>1%</td>
</tr>
<tr>
<td>Preparing for Standardized Tests</td>
<td>18%</td>
<td>52%</td>
<td>24%</td>
<td>3%</td>
<td>3%</td>
</tr>
</tbody>
</table>

10.3 The Chi-Square Test of Independence

Suppose you take a random sample of people from your community. From this sample alone, you could probably tell that being over 40 and having some gray hair are not independent characteristics in the population of all people. If your community is typical, people over 40 are more likely to have some gray hair than are people 40 or younger. The two categorical variables for this single sample are *age* (with categories over 40 and 40 or younger) and *hair* (with categories has some gray hair and doesn’t have gray hair). For each of these two variables, each member of the population falls into exactly one category.

When all you can observe is a single sample from a population, such as the sample from your community, you can use a chi-square test of independence to decide whether it is reasonable to believe that two different variables are independent in that population. In Activity 10.3a, you will collect some data that will be used to illustrate the ideas in this section.

**ACTIVITY 10.3a Independent or Not?**

1. In the next step, you will count the number of females and males in your class. You will also ask if the last digit of their phone number is even or odd. Do you think that the categorical variables *gender* and *even/odd phone number* are independent? Explain.

2. Collect the data described in step 1 and organize the data in a two-way table.

(continued)
3. Using the definition of independence in Chapter 5, determine if gender and even/odd phone number are independent variables in the selection of a student at random from your class.

4. Based on your answer to step 3, what are you willing to conclude about the independence of the two variables in the population of all students? Explain.

5. Even if gender and even/odd phone number are independent variables in the population of all students, why are you likely to find that this is not the case in your class?

6. In Activity 5.4a, you determined whether you are right-eye dominant or left-eye dominant. Retrieve the table you made then, or make another two-way table for your class in which one categorical variable is eye dominance (left or right) and one categorical variable is hand dominance (left or right).

7. For a randomly selected student from your class, are the events right-eye dominant and right-hand dominant independent according to the definition of independence in Chapter 5?

---

**The Chi-Square Test of Independence**

D18. In Chapter 5, you showed that two events $A$ and $B$ are independent by verifying that one of these statements is true (and if one statement is true, all three are true):

$$P(A) = P(A|B) \quad P(B) = P(B|A) \quad P(A \text{ and } B) = P(A) \cdot P(B)$$

Why do you need a test of independence here when you already have one from Chapter 5?

D19. Why do you think a chi-square test of independence is sometimes called a test of “association”?

---

**Tabular and Graphical Displays of the Data**

Now you will see what two-way tables and graphical displays of categorical data look like when two variables are dependent (or associated). In Chapter 5 you saw the data on the *Titanic* disaster, repeated in Display 10.52.

<table>
<thead>
<tr>
<th>Class of Travel</th>
<th>First</th>
<th>Second</th>
<th>Third</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Survived?</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>203</td>
<td>118</td>
<td>178</td>
<td>499</td>
</tr>
<tr>
<td>No</td>
<td>122</td>
<td>167</td>
<td>528</td>
<td>817</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>325</td>
<td>285</td>
<td>706</td>
<td>1316</td>
</tr>
</tbody>
</table>

Display 10.52 *Titanic* survival data. [Source: *Journal of Statistics Education* 3, no. 3 (1995).]
In testing for independence, there is only one population with each subject or unit categorized according to two variables. For the Titanic data, the categorical variables are Class of Travel and Survival Status. Recall from Section 10.2 that the number of first-class travelers who survived, 203, is a joint frequency, the number of survivors, 499, is a row marginal frequency and the number of second-class travelers, 285, is a column marginal frequency. Of great importance to the notion of independence are the conditional relative frequencies. The column conditional relative frequencies, shown in Display 10.53, give the proportions of those who survived or not for each class of traveler. Notice that these are quite different from column to column, indicating that the proportion of survivors changes dramatically from class to class. In other words, the survival rate depends on (or is associated with) the class of travel.

<table>
<thead>
<tr>
<th>Survived?</th>
<th>First</th>
<th>Second</th>
<th>Third</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.62</td>
<td>0.41</td>
<td>0.25</td>
</tr>
<tr>
<td>No</td>
<td>0.38</td>
<td>0.59</td>
<td>0.75</td>
</tr>
<tr>
<td>Total</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Display 10.53 Column conditional relative frequencies for the Titanic data.

Will the row conditional relative frequencies tell much the same story? Display 10.54 gives these data, which show what proportion of the survivors fall into each class of travel as well as what proportion of the non-survivors fall into each class of travel. Again, these conditional distributions are quite different, indicating that there is evidence of an association between survival status and class of travel.

<table>
<thead>
<tr>
<th>Survived?</th>
<th>First</th>
<th>Second</th>
<th>Third</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.41</td>
<td>0.23</td>
<td>0.36</td>
<td>1.00</td>
</tr>
<tr>
<td>No</td>
<td>0.15</td>
<td>0.20</td>
<td>0.65</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Display 10.54 Row conditional relative frequencies for the Titanic data.

The segmented bar chart was a natural graphic to use for studying homogeneity because, in that situation, you had different populations that could be compared with side-by-side bars. In studying independence you have one population and one sample categorized by two variables, so you can construct a segmented bar chart on either variable, as shown in Display 10.55 (on the next page). You can tell that the two variables are not independent by observing that the bars aren’t divided into segments according to the same proportions. For example, in the left-hand chart, the first bar is more than two-thirds “Survived,” whereas the third bar is only about one-third “Survived.”
The column chart in Display 10.56 treats the two variables symmetrically. You can see that the two variables do not appear to be independent because the columns in the back row don’t follow the same pattern as the columns in the front row. From left to right, the heights of the columns in the back row go shortest, middle, tallest; in the front row, the heights go tallest, shortest, middle.

**Display 10.55**  Segmented bar charts for the *Titanic* survival data.

**Display 10.56**  A column chart for the *Titanic* survival data.

**DISCUSSION**

**Tabular and Graphical Displays of the Data**

D20. Refer to the three plots in Displays 10.55 and 10.56.

a. For each plot, give one fact that the plot shows more clearly than the other two plots.

b. How can you tell from the right-hand segmented bar chart that *class of travel* and *survival status* do not appear to be independent?

D21. Construct two different segmented bar charts and one column chart to display the data your class collected in Activity 10.3a on the variables *gender* and *even/odd phone number*. How do your plots show whether these variables appear to be independent?

D22. What does a column chart look like if two variables are independent?
### Expected Frequencies in a Chi-Square Test of Independence

In a chi-square test of independence, the null hypothesis is that the variables are independent. As you learned in Chapter 5, events $A$ and $B$ are independent if and only if $P(A \text{ and } B) = P(A) \cdot P(B)$. With this definition, you can fill in the two-way table of expected frequencies in Display 10.57 if you assume that handedness and eye color are independent variables.

<table>
<thead>
<tr>
<th>Eye Color</th>
<th>Handedness</th>
<th>Right</th>
<th>Left</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>—?—</td>
<td>—?—</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Brown</td>
<td>—?—</td>
<td>—?—</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>24</td>
<td>6</td>
<td>30</td>
<td></td>
</tr>
</tbody>
</table>

Display 10.57 A possible two-way table for the variables handedness and eye color for $n = 30$.

Assuming independence, the probability that a randomly selected person is right-handed and blue-eyed is

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

$$P(\text{right-handed and blue-eyed}) = P(\text{right-handed}) \cdot P(\text{blue-eyed})$$

$$= \frac{24}{30} \cdot \frac{10}{30}$$

$$= \frac{4}{15}$$

The expected frequency of people who are right-handed and blue-eyed is then

$$P(\text{right-handed and blue-eyed}) \cdot (\text{grand total}) = \frac{4}{15}(30) = 8$$

In general, if two variables are independent, the probability that a randomly selected observation falls into the cell in column $C$ and row $R$ is

$$P(R \text{ and } C) = P(R) \cdot P(C) = \frac{\text{row } R \text{ total}}{\text{grand total}} \cdot \frac{\text{column } C \text{ total}}{\text{grand total}}$$

Thus, the expected number of observations that fall into this cell is

$$P(R \text{ and } C) \cdot (\text{grand total}) = \frac{\text{row } R \text{ total}}{\text{grand total}} \cdot \frac{\text{column } C \text{ total}}{\text{grand total}} \cdot (\text{grand total})$$

$$= \frac{\text{row } R \text{ total}}{\text{grand total}} \cdot \frac{\text{column } C \text{ total}}{\text{grand total}}$$

This is the same formula as in Section 10.2.
Frequencies in a Chi-Square Test of Independence

D23. Refer to your data from Activity 10.3a.

a. Complete an expected frequencies table for your class assuming gender and even/odd phone number are independent variables.

b. Complete an expected frequencies table for your class assuming handedness and eyedness are independent variables.

c. For which of the two tables in parts a and b are the expected and observed values farther apart?

Procedure for a Chi-Square Test of Independence

Now that you have reviewed the idea of independence, it’s time to organize all the steps needed when you examine a sample from a population in order to decide if two variables are independent in that population.

The steps in a chi-square test of independence are much the same as the steps in the chi-square test of homogeneity.

Chi-Square Test of Independence

1. **Name the test and check conditions.**
   - A simple random sample is taken from one large population.
   - Each outcome can be classified into one cell according to its category on one variable and its category on a second variable.
   - The expected frequency in each cell is 5 or greater.

2. **State the hypotheses.**
   
   \[ H_0: \text{The two variables are independent. That is, suppose one member is selected at random from the population. Then, for each cell, the probability that the member falls into both category } A \text{ and category } B, \text{ where } A \text{ is the category from the first variable and } B \text{ is the category from the second variable, is equal to } P(A) \cdot P(B). \]

   \[ H_1: \text{The two variables are not independent.} \]

3. **Compute the value of the test statistic, approximate the } P\text{-value, and draw a sketch.}**

   The test statistic is

   \[ \chi^2 = \sum \frac{(O - E)^2}{E} \]

   (continued)
For each cell in the table, \( O \) is the observed frequency and \( E \) is the expected frequency where:

\[
E = \frac{(\text{row total}) \cdot (\text{column total})}{\text{grand total}}
\]

The \( P \)-value is the probability of getting a value of \( \chi^2 \) as extreme as or more extreme than the one in the sample, assuming the null hypothesis is true. Get the \( P \)-value from your calculator or approximate the \( P \)-value by comparing the value of \( \chi^2 \) to the appropriate value of \( \chi^2 \) in Table C on page 827 with \( df = (r - 1)(c - 1) \), where \( r \) is the number of categories for one variable and \( c \) is the number of categories for the other variable.

4. **Write your conclusion linked to your computations and in the context of the problem.** If the \( P \)-value is smaller than \( \alpha \) (or, equivalently, if \( \chi^2 \) is larger than the critical value for the given \( \alpha \)), then reject the null hypothesis. If not, there is no evidence that the null hypothesis is false, so you do not reject it.

**Example: Scottish Children**

The information in Display 10.58 shows the eye color and hair color of 5387 Scottish children. You might suspect that these two categorical variables are not independent because people with darker hair colors tend to have darker eye colors, whereas people with lighter hair colors tend to have lighter eye colors. Test to see if there is an association between hair color and eye color.

<table>
<thead>
<tr>
<th>Hair Color</th>
<th>Fair</th>
<th>Red</th>
<th>Medium</th>
<th>Dark</th>
<th>Black</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>326</td>
<td>38</td>
<td>241</td>
<td>110</td>
<td>3</td>
</tr>
<tr>
<td>Light</td>
<td>688</td>
<td>116</td>
<td>584</td>
<td>188</td>
<td>4</td>
</tr>
<tr>
<td>Medium</td>
<td>343</td>
<td>84</td>
<td>909</td>
<td>412</td>
<td>26</td>
</tr>
<tr>
<td>Dark</td>
<td>98</td>
<td>48</td>
<td>403</td>
<td>681</td>
<td>85</td>
</tr>
</tbody>
</table>


**Solution**

From the column chart in Display 10.59, you can see that it does indeed appear to be the case that children with darker hair colors tend to have darker eye colors, whereas children with lighter hair colors tend to have lighter eye colors. Compare, especially, the rows for fair hair and dark hair.
This situation satisfies the conditions for a chi-square test of independence if the children can be considered a simple random sample taken from one large population. Each child in the sample falls into one hair color category and one eye color category. The Data Desk printout in Display 10.60 shows the expected frequencies under the assumption of independence. The expected frequency in each cell is 5 or greater.

The null hypothesis is

\[ H_0: \text{Eye color and hair color are independent.} \]

The alternative hypothesis is

\[ H_a: \text{Eye color and hair color are not independent.} \]

Your calculator gives a \( \chi^2 \) value of 1240, extremely far out in the tail, and a corresponding \( P \)-value of approximately 0. [See Calculator Note 10E to learn how to use your calculator to find these statistics.]

Display 10.60 also gives the computation of the test statistic, \( \chi^2 \), and the \( P \)-value. Notice that the expected frequencies in each cell are computed from the marginal frequencies. Display 10.60 also includes the conditional relative frequencies for the columns (hair color): Among those with fair hair, 22.4% have blue eyes; among those with black hair, only 2.54% have blue eyes.
Rows are levels of **Eye Color**: Count  
Columns are levels of **Hair Color**: Count  
No Selector  

<table>
<thead>
<tr>
<th></th>
<th>Fair</th>
<th>Red</th>
<th>Medium</th>
<th>Dark</th>
<th>Black</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>326</td>
<td>38</td>
<td>241</td>
<td>110</td>
<td>3</td>
<td>718</td>
</tr>
<tr>
<td></td>
<td>22.4</td>
<td>13.3</td>
<td>11.3</td>
<td>7.91</td>
<td>2.54</td>
<td>13.3</td>
</tr>
<tr>
<td></td>
<td>193.928</td>
<td>38.1192</td>
<td>284.828</td>
<td>185.398</td>
<td>15.7275</td>
<td>718</td>
</tr>
<tr>
<td>Light</td>
<td>688</td>
<td>116</td>
<td>584</td>
<td>188</td>
<td>4</td>
<td>1580</td>
</tr>
<tr>
<td></td>
<td>47.3</td>
<td>40.6</td>
<td>27.3</td>
<td>13.5</td>
<td>3.39</td>
<td>29.3</td>
</tr>
<tr>
<td></td>
<td>426.750</td>
<td>83.8834</td>
<td>626.779</td>
<td>407.978</td>
<td>34.6092</td>
<td>1580</td>
</tr>
<tr>
<td>Medium</td>
<td>343</td>
<td>84</td>
<td>909</td>
<td>412</td>
<td>26</td>
<td>1774</td>
</tr>
<tr>
<td></td>
<td>23.6</td>
<td>29.4</td>
<td>42.5</td>
<td>29.6</td>
<td>22.0</td>
<td>32.9</td>
</tr>
<tr>
<td></td>
<td>479.148</td>
<td>94.1830</td>
<td>703.738</td>
<td>458.072</td>
<td>38.8587</td>
<td>1774</td>
</tr>
<tr>
<td>Dark</td>
<td>98</td>
<td>48</td>
<td>403</td>
<td>681</td>
<td>85</td>
<td>1315</td>
</tr>
<tr>
<td></td>
<td>6.74</td>
<td>16.8</td>
<td>18.9</td>
<td>49.0</td>
<td>72.0</td>
<td>24.4</td>
</tr>
<tr>
<td></td>
<td>355.174</td>
<td>69.8144</td>
<td>521.655</td>
<td>339.552</td>
<td>28.8045</td>
<td>1315</td>
</tr>
<tr>
<td>Total</td>
<td>1455</td>
<td>286</td>
<td>2137</td>
<td>1391</td>
<td>118</td>
<td>5387</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>5387</td>
</tr>
<tr>
<td></td>
<td>1455</td>
<td>286</td>
<td>2137</td>
<td>1391</td>
<td>118</td>
<td>5387</td>
</tr>
</tbody>
</table>

Table contents:  
Count  
Percent of Column Total  
Expected Values  
Chi-square = 1240 with 12 df  
p ≤ 0.0001  

**Display 10.60** Data Desk printout of a chi-square test of independence for the Scottish children example. 

A sketch of the distribution is shown in Display 10.61. 

![Chi-square distribution with 12 degrees of freedom](image)

**Display 10.61** Chi-square distribution with 12 degrees of freedom. 

Reject the null hypothesis. These are not results you would expect for a sample from a population in which there is no association between eye color and hair color. As you can see from Display 10.61, a $\chi^2$ value of 1240 is much larger than the value of 32.91 that cuts off an upper tail of 0.001. Thus, a value of $\chi^2$ of 1240 or larger is extremely unlikely to occur in a sample of this size if hair color and eye color are independent. Examining the table and the column chart, you might conclude that darker eye colors tend to go with darker hair colors and lighter eye colors tend to go with lighter hair colors.
Procedures for a Chi-Square Test of Independence

D24. Refer to your data from Activity 10.3a.

a. Assuming your class is a random sample of students, test the hypothesis that gender and even or odd phone number are independent characteristics among students.

b. Assuming your class is a random sample of students, test the hypothesis that handedness and eyedness are independent characteristics among students.

Homogeneity Versus Independence

Ms. C: Did you notice that a chi-square test of homogeneity and a chi-square test of independence are performed in exactly the same way, except for the wording of the hypotheses and the conclusion? The two null hypotheses say exactly the same thing about the table: The columns are proportional.

Class: So how do we know which test to use?

Ms. C: You can’t tell from the table itself. You have to find out how the data were collected. Did they take one sample of a fixed size from one population and then classify each person according to two categorical variables? If so, it’s a test of independence. Or did they sample separately from two or more populations and then classify each person according to one categorical variable? If so, it’s a test of homogeneity.

Class: Huh?

Ms. C: It’s simple. Suppose you want to find out whether there is an association between age and wearing blue jeans. You take a random sample of 50 people under age 40. Then you take a random sample of 50 people age 40 or older. Your table might look like Display 10.62.

<table>
<thead>
<tr>
<th>Age</th>
<th>Under 40</th>
<th>40 or Older</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wears Blue Jeans?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Now</td>
<td>21</td>
<td>20</td>
</tr>
<tr>
<td>Not Now, but Sometimes</td>
<td>23</td>
<td>25</td>
</tr>
<tr>
<td>Never</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

Display 10.62 Sample table for blue jeans data, collected from two populations.

Class: We get it! We have to do a test of homogeneity because we sampled separately from two populations: people under 40 and older people. There are two populations and one variable about jeans.

Jodain: Now I see what you mean by the columns being proportional. For people under 40, 42% are wearing jeans now, 46% are not
wearing them now, and 12% never wear them. That’s about the same distribution as for people 40 or older, so we won’t be able to reject the hypothesis that these populations are homogeneous with respect to wearing jeans. Jeans-wearing behavior is about the same for both age groups.

**Ms. C:** Right. In the homogeneity case, you have one variable but several populations. The column—or sometimes the row—proportions represent the separate distributions for those populations. Now suppose you go out and take a random sample of 100 people, classify them, and get the data in Display 10.63. This time you sampled from just one population.

<table>
<thead>
<tr>
<th>Wears Blue Jeans?</th>
<th>Age</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Under 40</td>
<td>40 or Older</td>
<td>Total</td>
</tr>
<tr>
<td>Now</td>
<td>23</td>
<td>17</td>
<td>40</td>
</tr>
<tr>
<td>Not Now, but Sometimes</td>
<td>28</td>
<td>22</td>
<td>50</td>
</tr>
<tr>
<td>Never</td>
<td>6</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>57</td>
<td>43</td>
<td>100</td>
</tr>
</tbody>
</table>

**Display 10.63** Sample table for blue jeans data, collected from one population.

**Class:** We use a test of independence because that’s the only other possibility.

**Jodain:** We use a test of independence because there is only one sample, from the population of all people. We categorize each person in the sample on the two variables, age and jeans-wearing behavior. The columns are roughly proportional, so we’ll conclude that we can’t reject the hypothesis that the two variables are independent. Jeans-wearing behavior and age aren’t associated. But this time we didn’t predetermine how many people would be in each age group.

**Ms. C:** Right again, Jodain.

**Class:** Jodain is always right!

**Ms. C:** Also, it now makes sense to talk about conditional distributions for the rows. You can estimate that \( \frac{23}{40} \), or about 58%, of people now wearing jeans are under age 40 and about 42% are age 40 or older. That statement wouldn’t make sense in the homogeneity case.

**Jodain:** Then why wouldn’t we always design our study as a test of independence? Then we can look at the table both ways, and as a bonus we get an estimate of the percentage of people in each age group!

**Ms. C:** It depends on what you want to find out. Suppose your research hypothesis is that the proportion of people who wear jeans differs for the two age groups. You would design the study as a test of homogeneity, taking an equal sample size from each age group.
That gives you the best chance of rejecting a false null hypothesis. In short, if you want to compare populations, take independent random samples from each and conduct a test of homogeneity. If you want to describe association between two variables in a single population, take a single random sample and conduct a test of independence.

Class: How old are you, Ms. C?

**Homogeneity Versus Independence**

D25. Suppose you want to study whether people with straight hair and people with curly hair are equally likely to use blow-dryers. You are considering two designs for your survey:

I. You take a random sample of 50 people with straight hair and ask them whether they use a blow-dryer. You take a random sample of 50 people with curly hair and ask them whether they use a blow-dryer.

II. You take a random sample of 100 people, note if they have curly hair or straight hair, and ask them whether they use a blow-dryer.

a. Which of the two designs would result in a test of homogeneity, and which would result in a test of independence?

b. How are the data from the samples likely to be different?

c. Is there any reason why one design would be better than the other?

d. Which of the two designs could also be correctly analyzed by using a z-test for the difference between two proportions?

**Strength of Association and Sample Size**

A chi-square test of independence tells you if there is evidence of an association between two categorical variables. As is typical of a test of significance, it does not tell you about the size of any difference in the proportions from column to column (the **strength of the association**) or about whether these differences are of practical importance. As an illustration, Display 10.64 shows three separate samples of responses on the same two variables.

<table>
<thead>
<tr>
<th>Response</th>
<th>Sample 1 A</th>
<th>Sample 1 B</th>
<th>Sample 2 A</th>
<th>Sample 2 B</th>
<th>Sample 3 A</th>
<th>Sample 3 B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>48</td>
<td>52</td>
<td>96</td>
<td>104</td>
<td>960</td>
<td>1040</td>
</tr>
<tr>
<td>No</td>
<td>52</td>
<td>48</td>
<td>104</td>
<td>96</td>
<td>1040</td>
<td>960</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>100</td>
<td>200</td>
<td>200</td>
<td>2000</td>
<td>2000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample 1 A B</th>
<th>Sample 2 A B</th>
<th>Sample 3 A B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difference between proportions of A and B saying yes</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>0.32</td>
<td>0.64</td>
</tr>
<tr>
<td>$P$-value</td>
<td>0.57</td>
<td>0.42</td>
</tr>
</tbody>
</table>

**Display 10.64**  Tables showing how sample size can affect the test statistic.
In each sample, the columns are almost proportional—the difference between the proportions in the first and second columns is only 0.04. This is not a very strong association; in most applications, a difference of 0.04 between two proportions would be of little practical significance. However, if you do a chi-square test of independence on each of the three samples, you will see that as the sample size increases, the test statistic gets larger and the \( P \)-value gets smaller. At a sample size of 4000, the association between the variables becomes highly significant statistically, but the strength of the association is no larger than for the smaller samples. For fixed conditional distributions, the numerical value of the \( \chi^2 \) statistic is directly proportional to the sample size.

Two bits of advice are in order: Be wary when using this test (or any test) with very large data sets, and always compute and compare proportions to see if the statistical significance has any practical value.

**Strength of Association and Sample Size**

D26. As in all hypothesis tests, the \( P \)-value measures the strength of the evidence against the null hypothesis (or, in favor of the alternative hypothesis) provided by the data. If the \( P \)-value is very small, there is much evidence in favor of the alternative hypothesis.

a. In the context of a chi-square test of independence, discuss the difference between the statements “There is much evidence in favor of association” and “There is evidence of a strong association.”

b. Suppose you have three \( 2 \times 2 \) tables of frequencies for two categorical variables and a chi-square test of independence gives these results:

   - Table A: very small \( P \)-value and a small difference between conditional proportions
   - Table B: very small \( P \)-value and a large difference between conditional proportions
   - Table C: very large \( P \)-value and a small difference between conditional proportions

   For Table A, is there much or little evidence of association? Is there evidence of strong association or evidence of weak association? Answer these questions for Table B and Table C.

**Summary 10.3: The Chi-Square Test of Independence**

Suppose there is no reason to suspect anything other than complete independence between two variables in a population. It would be a near-miracle to get a random sample in which the variables satisfy the definition of independence in Chapter 5. Requiring \( P(A \text{ and } B) \) to be exactly equal to \( P(A) \cdot P(B) \) for every cell of the table is too rigid a position to take when dealing with random samples. A chi-square test of independence is much more useful: Is it reasonable to assume that the random
sample came from a population in which the two variables are independent? Use the chi-square test of independence when

- a simple random sample of a fixed size is taken from one large population
- each outcome can be classified into one cell according to its category on one variable and its category on a second variable
- you want to know if it's plausible that this sample came from a population in which these two categorical variables are independent

In a chi-square test of independence, the expected frequencies are calculated from the sample data:

\[
E = \frac{(\text{row total}) \cdot (\text{column total})}{\text{grand total}}
\]

Each value of \(E\) should be 5 or greater.

The number of degrees of freedom for a two-way table with \(r\) rows and \(c\) columns is

\[df = (r - 1)(c - 1)\]

If the result from your sample is very different from the result expected under independence, then \(\chi^2\) will be large and you will reject the hypothesis that the variables are independent. However, even if the test tells you there is evidence of an association (dependence), it does not tell you anything about the strength of the association or whether it is of practical importance.

### Practice

#### The Chi-Square Test of Independence

P22. Which pairs of variables do you believe are independent in the population of U.S. students? Explain.

A. hair color and eye color  
B. type of music preferred and ethnicity  
C. gender and color of shirt  
D. type of movie preferred and gender  
E. eye color and class year  
F. class year and whether taking statistics

#### Tabular and Graphical Display of Data

P23. Which column charts in Display 10.65 display variables that are independent?

A.  
B.  
C.  
D.  

Display 10.65  Four column charts.
Expected Frequencies in a Chi-Square Test of Independence

P24. Display 10.66 shows the marginal totals for a sample of 2237 residents of the United States classified according to gender and handedness. On a copy of the table, fill in the expected frequencies for each cell under the hypothesis that gender and handedness are independent variables. Do you think these variables really are independent?

<table>
<thead>
<tr>
<th>Gender</th>
<th>Right-Handed</th>
<th>Left-Handed</th>
<th>Ambidextrous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>2004</td>
<td>205</td>
<td>28</td>
</tr>
<tr>
<td>Women</td>
<td>1170</td>
<td>1067</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2237</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Display 10.66 Table of expected frequencies for a test of independence between gender and handedness.

Procedure for a Chi-Square Test of Independence

P25. Is there a difference between handedness patterns in men and women? A good set of data to help you answer this question comes from the government's 5-year Health and Nutrition Survey (HANES) of 1976–80, which recorded the gender and handedness of a random sample of 2237 individuals from across the country. The observed frequencies for men and women in each of three handedness categories are shown in Display 10.67. Use an appropriate statistical test to answer the question posed. If the patterns differ significantly, explain where the main differences occur.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Right-Handed</th>
<th>Left-Handed</th>
<th>Ambidextrous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>934</td>
<td>113</td>
<td>20</td>
</tr>
<tr>
<td>Women</td>
<td>1070</td>
<td>92</td>
<td>8</td>
</tr>
<tr>
<td>Total</td>
<td>1067</td>
<td>1170</td>
<td></td>
</tr>
</tbody>
</table>


Homogeneity Versus Independence

P26. A Time magazine poll asked a random sample of people from across America this question: “How much effort are you making to eat a healthy and nutritionally balanced diet?” The results are given in Display 10.68.

<table>
<thead>
<tr>
<th>Effort</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very Serious</td>
<td>31%</td>
<td>43%</td>
</tr>
<tr>
<td>Somewhat Serious</td>
<td>47%</td>
<td>43%</td>
</tr>
<tr>
<td>Not Very Serious</td>
<td>12%</td>
<td>9%</td>
</tr>
<tr>
<td>Don’t Really Try</td>
<td>10%</td>
<td>5%</td>
</tr>
</tbody>
</table>

Display 10.68 Results of a poll on eating habits. [Source: Time, January 30, 1995.]

a. Explain what the percentages measure. Given the way the data were collected, would you do a test of homogeneity or a test of independence?
b. Suppose the sample size was 1000 and it just happened to be equally split between men and women. Is there evidence of dependence between gender and an effort toward a healthy diet?
c. Suppose the sample size was only 500 and it just happened to be equally split between men and women. Is there evidence of dependence between gender and an effort to eat a healthy diet?

P27. A health science teacher had his class do a sample survey of the students at a nearby middle school. Two of the questions were “Do you eat breakfast at least three times a week?” and “Do you think your diet is healthy?” The responses are shown in Display 10.69.

Do the answers to these two questions appear to be independent, based on these data? Conduct a statistical test.

<table>
<thead>
<tr>
<th>Healthy Diet?</th>
<th>Breakfast at Least Three Times a Week</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Display 10.69 Table of responses from middle school students regarding eating habits.
Strength of Association and Sample Size

P28. The American Community Survey is a nationwide survey that collects and analyzes data at the community level. Eventually it will replace the long form used in the census conducted every 10 years. In the 2004 survey, about 322,400 interviews were conducted with adults over age 25.

a. In that 2004 survey, 80.1% of the men had at least a high school education, and 80.7% of the women had at least a high school education. A two-way table for the results is given in Display 10.70. Is there a significant association between gender of the respondent and attainment of a high school education? If so, is this evidence of a strong association, or strong evidence of a weak association?

<table>
<thead>
<tr>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>High School Education or More</td>
<td>123,486</td>
</tr>
<tr>
<td>Less Than High School Education</td>
<td>30,679</td>
</tr>
<tr>
<td>Total</td>
<td>154,165</td>
</tr>
</tbody>
</table>

Display 10.70 Results of a survey regarding the education level of 322,400 adults in 2004. [Source: www.census.gov.]

b. In the same survey, 26.1% of the men had a bachelor’s degree or higher, while 22.8% of the women had a bachelor’s degree or higher. Is the association between college degree and gender stronger than that between high school degree and gender? Explain your reasoning.

c. In large sample surveys like the American Community Survey, almost any small difference can show up as being statistically significant but not necessarily of practical importance. Would you call either of the differences seen in parts a and b practically important? Why or why not?

Exercises

Always include all steps when doing a statistical test.

E33. According to a National Center for Education Statistics report, 48,574,000 children were projected to be enrolled in the nation’s K–12 public schools in 2006. About 69% were projected to be enrolled in grades K–8 and 31% in grades 9–12. It was also projected that about 16.8% of the students would be in the Northeast, 22.1% in the Midwest, 36.5% in the South, and 24.6% in the West. [Source: nces.ed.gov.]

a. For what reason might the variables region of country and grade level be associated?

b. Construct a two-way table showing the proportion of students who fall into each cell under the assumption of independence.

c. Construct a two-way table showing the number of students who fall into each cell under the assumption of independence.

E34. The first U.S. National Health and Nutrition Examination Survey in the 1980s reported on the age of a mother at the birth of her first child and whether the mother eventually developed breast cancer. Of the 6168 mothers in the sample, 26.39% had their first child at age 25 or older. Of the 6168 mothers in the sample, 98.44% had not developed breast cancer. [Source: Jessica Utts, Seeing Through Statistics, 2d ed. (Pacific Grove, Calif.: Duxbury, 1999), p. 209.]

a. For what reason might the variables age at birth of first child and whether developed breast cancer be associated?

b. Construct a two-way table showing the percentage of mothers who fall into each cell under the assumption of independence.

c. Construct a two-way table showing the number of mothers who fall into each cell under the assumption of independence.

E35. A student surveyed a random sample of 300 students in her large college and collected the data in Display 10.71 on the variables class year and favorite team sport.
E36. In combined polls of 3026 adults taken over a 3-year period, the Gallup Organization got the results in Display 10.72 regarding the level of exercise attained regularly by adult men and women. (Assume the respondents are equally split between men and women.)

<table>
<thead>
<tr>
<th></th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>33%</td>
<td>26%</td>
</tr>
<tr>
<td>Medium</td>
<td>23%</td>
<td>16%</td>
</tr>
<tr>
<td>Low</td>
<td>20%</td>
<td>26%</td>
</tr>
<tr>
<td>Sedentary</td>
<td>24%</td>
<td>32%</td>
</tr>
</tbody>
</table>

Display 10.72 Results of a survey regarding the exercise level of 3026 adults, taken over a 3-year period. [Source: poll.gallup.com, 2005.]

a. Change the percentages into observed frequencies and perform a chi-square test to determine if exercise and gender are independent.

b. Could you have made a Type I error?

c. Describe how the survey should have been designed in order to test for homogeneity of male and female populations with regard to level of exercise.

E37. Display 10.73 is an example of a single population, Titanic passengers, that has been sorted by two variables: gender and survival status. These data cannot reasonably be considered a random sample from any well-defined population.

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survived?</td>
<td>367</td>
<td>344</td>
<td>711</td>
</tr>
<tr>
<td>Yes</td>
<td>1364</td>
<td>126</td>
<td>1490</td>
</tr>
<tr>
<td>No</td>
<td>1731</td>
<td>470</td>
<td>2201</td>
</tr>
</tbody>
</table>

Display 10.73 Titanic passengers sorted by gender and survival status.

[Source: www.utahcrossroads.org, June 2002.]

da. Construct a plot that displays these data to see whether the variables gender and survival status appear to be independent.

db. Test to see if the association between the variables gender and survival status can reasonably be attributed to chance or if you should look for some other explanation.

dc. The fate of the members of the Donner party, who were trapped in the Sierra Nevada mountain range over the winter of 1846–47, is shown in Display 10.74. These data cannot reasonably be considered a random sample from any well-defined population.

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survived?</td>
<td>23</td>
<td>25</td>
</tr>
<tr>
<td>Yes</td>
<td>33</td>
<td>9</td>
</tr>
</tbody>
</table>

Display 10.74 The Donner party survival data. [Source: www.utahcrossroads.org, June 2002.]

10.3 The Chi-Square Test of Independence 727
a. Construct a plot that displays these data to see whether the variables gender and survival status appear to be independent.

b. Test to see if the association between the variables gender and survival status can reasonably be attributed to chance or if you should look for some other explanation.

E39. Are people involved in alcohol-related accidents on public roadways in the United States just as likely to survive on Super Bowl Sunday as on other Sundays? To help answer this question, investigators looked at data for alcohol-related crashes during the 4 hours after the telecast of the first 27 Super Bowls. They compared these data with the data for alcohol-related crashes during the same time period on the Sundays the week before and immediately after the Super Bowl Sundays. Display 10.75 gives the number of people who were killed and who survived alcohol-related crashes for the 27 Super Bowl Sundays and the 54 control Sundays.

<table>
<thead>
<tr>
<th></th>
<th>Super Bowl Sundays</th>
<th>Control Sundays</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Killed</td>
<td>284</td>
<td>163</td>
<td>447</td>
</tr>
<tr>
<td>Survived</td>
<td>48</td>
<td>218</td>
<td>266</td>
</tr>
<tr>
<td>Total</td>
<td>332</td>
<td>381</td>
<td>713</td>
</tr>
</tbody>
</table>


a. Is this an experiment, a sample survey, or an observational study?

b. Does the comparison of the survival rates on the two types of Sundays call for a test of homogeneity or a test of independence? Defend your choice.

c. Is there a statistically significant association between the type of Sunday and whether a person survived an alcohol-related accident? State your conclusion carefully, in light of parts a and b.

E40. Display 10.76 gives the smoking behavior by occupation type for a sample of white males in 1976. Assuming this can be considered a random sample of all white males in the United States in 1976, perform a chi-square test of independence to see if it is reasonable to assume that the variables occupation type and smoking behavior are independent.

<table>
<thead>
<tr>
<th></th>
<th>Blue Collar</th>
<th>Professional</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker?</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>43.2%</td>
<td>6.3%</td>
<td>50.5%</td>
<td>13,112</td>
</tr>
<tr>
<td>Former</td>
<td>30.9%</td>
<td>11.0%</td>
<td>58.1%</td>
<td>8,509</td>
</tr>
<tr>
<td>Never</td>
<td>30.5%</td>
<td>11.8%</td>
<td>58.7%</td>
<td>9,694</td>
</tr>
</tbody>
</table>


Chapter Summary
In this chapter, you have learned about three chi-square tests: a test of goodness of fit, a test of homogeneity, and a test of independence. Although each test is conducted in exactly the same way, the questions they answer are different.

In a chi-square goodness-of-fit test, you ask “Does this look like a random sample from a population in which the proportions that fall into these categories are the same as the proportions hypothesized?” This test is an extension of the test of a single proportion developed for the binomial case.

In a chi-square test of homogeneity, you ask “Do these samples from different populations look like random samples from populations in which the proportions
that fall into these categories are equal?" This test is an extension of the test for the equality of two binomial proportions.

In a chi-square test of independence, you ask “Does this sample look like a random sample from a population in which these two categorical variables are independent (not associated)?” This test is not equivalent to any test developed earlier in this book. It is, however, related to the concept of independent events and the computation of \( P(A \text{ and } B) \).

**Review Exercises**

E41. This question was asked of random samples of about 1000 residents of the United States in each of a succession of years: “Which of the following statements reflects your view of when the effects of global warming will begin to happen?” The percentages selecting each choice, by year, are given in Display 10.77 (at the bottom of the page). Is there evidence of change in the pattern in these percentages across the years? If so, describe the pattern of change you see.

E42. “Women are more critical of environmental conditions than are men.” So says a Gallup Poll report on a survey of 1004 adults taken in 2005. Results are shown in Display 10.78. The percentages in the first row of data show that 31% of men reported a positive view of current environmental conditions whereas only 18% of women reported such a view. Assume that there were approximately equal numbers of men and women.

a. Construct a plot that displays these data.

b. What test should you use to decide whether the data support the quoted claim?

c. Perform the test you think is best.

<table>
<thead>
<tr>
<th></th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>31%</td>
<td>18%</td>
</tr>
<tr>
<td>Mixed</td>
<td>23%</td>
<td>22%</td>
</tr>
<tr>
<td>Negative</td>
<td>44%</td>
<td>57%</td>
</tr>
<tr>
<td>Undesignated</td>
<td>2%</td>
<td>3%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Display 10.78 Results of a poll of 1004 adults regarding how critical they are of environmental conditions. [Source: poll.gallup.com, 2005.]

E43. Many years ago, Smith College, a residential college, switched to an unusual academic schedule that made it fairly easy for students to take most or all of their classes on the first three days of the week. (Smith has long since abandoned this experiment.) At the time, the infirmary staff wanted to know about after-hours use of the infirmary under this schedule. They gathered data for an entire academic year, recording the time and day of the week of each after-hours visit, along with the nature of the problem,
which they later classified as belonging to one of four categories: A visit to the infirmary was unnecessary for the problem, the problem required a nurse’s attention, the problem required a doctor’s attention, the problem required admission to the infirmary or to the local hospital. The results appear in Display 10.79.

<table>
<thead>
<tr>
<th>Nature of Problem</th>
<th>Mon</th>
<th>Tues</th>
<th>Wed</th>
<th>Thurs</th>
<th>Fri</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visit unnecessary</td>
<td>77</td>
<td>67</td>
<td>77</td>
<td>78</td>
<td>70</td>
</tr>
<tr>
<td>Nurse needed</td>
<td>80</td>
<td>66</td>
<td>53</td>
<td>73</td>
<td>62</td>
</tr>
<tr>
<td>Doctor needed</td>
<td>90</td>
<td>71</td>
<td>76</td>
<td>95</td>
<td>75</td>
</tr>
<tr>
<td>Admitted</td>
<td>61</td>
<td>49</td>
<td>42</td>
<td>52</td>
<td>28</td>
</tr>
</tbody>
</table>


a. Construct a plot to display these data.
b. Are there any interesting trends in the table? What might explain them?
c. Could you use a test of independence on these data? A test of homogeneity? Does the design of the study fit the conditions of these tests?
d. Perform the test you think is best, giving any necessary cautions about your conclusion.

e. How might you group the data to see whether the data for the weekdays when students attend class differ from the data for the weekdays when students don’t attend class?
f. Perform a chi-square test on your regrouped data.

E44. A 2002 Gallup poll asked these questions of a randomly selected national sample of approximately 1000 adults, 18 years and older: “Do you think that global warming will pose a serious threat to you or your way of life in your lifetime? Do you think that global warming will pose a serious threat to your children or the next generation of Americans in their lifetime?” The results are shown in Display 10.80.

<table>
<thead>
<tr>
<th>Response</th>
<th>Your Lifetime</th>
<th>Your Children’s Lifetime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>31%</td>
<td>35%</td>
</tr>
<tr>
<td>No</td>
<td>66%</td>
<td>57%</td>
</tr>
<tr>
<td>No Opinion</td>
<td>3%</td>
<td>8%</td>
</tr>
<tr>
<td>Total</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Display 10.80 Results of Gallup poll on the threat of global warming. [Source: www.americans-world.org, June 2002.]

a. Does this situation call for a chi-square test of homogeneity, a chi-square test of independence, or neither?
b. Perform the test you selected, or explain why neither test is appropriate.

E45. Vehicles can turn right, turn left, or continue straight ahead at a given intersection. It is hypothesized that vehicles entering the intersection from the south will continue straight ahead 50% of the time and that a vehicle not continuing straight ahead is just as likely to turn left as to turn right. A sample of 50 vehicles gave the counts shown in Display 10.81. Are the data consistent with the hypothesis?

<table>
<thead>
<tr>
<th>Number of vehicles</th>
<th>Straight</th>
<th>Left Turn</th>
<th>Right Turn</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>28</td>
<td>12</td>
<td>10</td>
</tr>
</tbody>
</table>

Display 10.81 Vehicle turning data.

E46. In testing a random-digit generator, one criterion is to be sure that each digit occurs \( \frac{1}{10} \) of the time.

a. Generate 100 random digits on your calculator, and use a chi-square test to see if it’s reasonable to assume that this criterion is met.
b. This criterion isn’t sufficient. Give an example of a sequence of digits that clearly is not random but in which each digit occurs \( \frac{1}{10} \) of the time. How might you test against this possibility?

E47. In a random sample of 100 college graduates, 100 high school graduates, and 100 people
who didn’t graduate from high school, 40% of the college graduates watched the last Super Bowl, as did 49% of the high school graduates and 37% of the people who didn’t graduate from high school. Is there a relationship between watching the Super Bowl and a person’s educational level?

a. Does the design of this study suggest a test of goodness of fit, homogeneity, or independence?

b. Perform the test you selected in part a.

E48. The data in Display 10.82 show how a random sample of people describe their political ideology and whether they usually think of themselves as Republicans, Independents, or Democrats. Does political ideology appear to be independent of party affiliation? If not, explain something of the nature of the dependency.

<table>
<thead>
<tr>
<th>Political Ideology</th>
<th>Liberal</th>
<th>Moderate</th>
<th>Conservative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democrat</td>
<td>313</td>
<td>387</td>
<td>174</td>
</tr>
<tr>
<td>Independent</td>
<td>287</td>
<td>469</td>
<td>292</td>
</tr>
<tr>
<td>Republican</td>
<td>79</td>
<td>185</td>
<td>407</td>
</tr>
</tbody>
</table>

Display 10.82 Results from the General Social Survey of the United States for the year 2000. [Source: www.icpsr.umich.edu, June 2002.]

E49. Display 10.83 shows a two-way table for a sample of size 180.

<table>
<thead>
<tr>
<th>Tree Type of Its Nearest Neighbor</th>
<th>I</th>
<th>II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree Type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>II</td>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>

a. Which cell counts, among A, B, C, and D, would be large if segregation of species is high? Would this lead to a large or small value of the $\chi^2$ statistic in a test of independence?

b. Which cell counts, among A, B, C, and D, would be large if segregation of species is low? Would this lead to a large or small value of the $\chi^2$ statistic in a test of independence?

E50. Foresters often are interested in the patterns of trees in a forest. For example, they gauge the integration or segregation of two species of trees (say, I and II) by looking at a sample of one species and then observing the species of its nearest neighbor. Data from such a study can be arrayed in a two-way table such as this one.

a. Enter a set of possible observed frequencies into a copy of the table. Pick observed values so that the variables clearly aren’t independent.

b. Can you enter a set of possible observed frequencies, different from the expected frequencies, into a copy of the table such that the variables clearly are independent?

E51. Display 10.84 (on the next page) gives information about the scores on the 1996 and 2001 AP Calculus AB Exams. The data for 1996 were obtained from all students who took the exam. The data for 2001 were obtained from a random sample of 200 students.

a. Suppose you want to investigate whether these data provide evidence of a change in the distribution of grades from 1996 to 2001. Is a chi-square test appropriate? If so, perform the test you selected, showing all steps. If not, explain why not.
b. Suppose you want to investigate whether these data provide evidence of an increase in the mean grade from 1996 to 2001. What test is appropriate? Perform this test, showing all steps.

e52. A common Gallup poll question often asked of a sample of residents of the United States is this: “In general, are you satisfied or dissatisfied with the way things are going in the United States at this time?” Display 10.85 shows the results of five such polls conducted from 2001 to 2005.

<table>
<thead>
<tr>
<th>Date of Poll</th>
<th>Percentage Satisfied</th>
<th>Percentage Dissatisfied</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>40</td>
<td>58</td>
</tr>
<tr>
<td>2004</td>
<td>43</td>
<td>55</td>
</tr>
<tr>
<td>2003</td>
<td>46</td>
<td>53</td>
</tr>
<tr>
<td>2002</td>
<td>52</td>
<td>44</td>
</tr>
<tr>
<td>2001</td>
<td>55</td>
<td>42</td>
</tr>
</tbody>
</table>

Display 10.85 Results of five polls taken from 2001 to 2005. [Source: poll.gallup.com, 2005.]

a. Why don’t the row percentages sum to 1?

b. Suppose each poll had sampled about 1000 residents (about the size of most Gallup polls). Is there evidence that the level of satisfaction changed significantly across these years? If so, describe the pattern of change.

e53. The data in Display 10.86 show the Titanic passengers sorted by two variables: class of travel and survival status. These data cannot reasonably be considered a random sample. They are the population itself, so you must be careful in stating the hypotheses and the conclusion for any test of significance. Test to see if the apparent lack of independence between the variables class of travel and survival status can reasonably be attributed to chance or if you should look for some other explanation. If the latter, what is that explanation?

<table>
<thead>
<tr>
<th>Class of Travel</th>
<th>Survived?</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First</td>
<td>203</td>
<td>122</td>
</tr>
<tr>
<td></td>
<td>Second</td>
<td>118</td>
<td>167</td>
</tr>
<tr>
<td></td>
<td>Third</td>
<td>178</td>
<td>928</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>499</td>
<td>817</td>
</tr>
</tbody>
</table>

Display 10.86 Titanic passengers sorted by class of travel and survival status.
E54. In 1988, there were 540 spouse-murder cases in the 75 largest counties in the United States. Display 10.87 gives the outcomes.

<table>
<thead>
<tr>
<th>Defendant</th>
<th>Husband</th>
<th>Wife</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not prosecuted</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>Plead guilty</td>
<td>146</td>
<td>87</td>
</tr>
<tr>
<td>Convicted at trial</td>
<td>130</td>
<td>69</td>
</tr>
<tr>
<td>Acquitted at trial</td>
<td>7</td>
<td>31</td>
</tr>
</tbody>
</table>

Display 10.87  Outcomes for spouse-murder defendants in large urban counties.
[Source: Bureau of Justice Statistics, Spouse Murder Defendants in Large Urban Counties, Executive Summary, NCJ-156831 (September 1995).]

a. What are the populations?
b. Construct a suitable plot to display the data. Describe what the plot shows.
c. Write suitable hypotheses and perform a chi-square test.

E55. In this chapter, you have seen three chi-square tests. Match each test (a–c) with its description (A–D) and its design (I–IV). (One description and one design will be left over.)

a. chi-square test of goodness of fit
b. chi-square test of homogeneity
c. chi-square test of independence

description:

A. You test that two populations have equal proportions of members that fall into each of a given set of categories.
B. You test that a population has the same proportion that falls into each of a given set of categories as some hypothesized distribution.
C. You test that you can predict the result for one categorical variable better if you know the result for the other.
D. You test that two populations are independent.

design:

I. one sample from one population sorted according to one categorical variable
II. one sample from one population sorted according to two categorical variables
III. two samples from two populations sorted according to one categorical variable
IV. two samples from two populations sorted according to two categorical variables
AP1. This partially completed table of expected frequencies is for a test of the independence of a father's handedness and his oldest child's handedness. What is the expected frequency in the cell marked “—?—”?

<table>
<thead>
<tr>
<th>Father</th>
<th>Left</th>
<th>Right</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left</td>
<td>64</td>
<td>72</td>
<td>136</td>
</tr>
<tr>
<td>Right</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Total</td>
<td>64</td>
<td>64</td>
<td>128</td>
</tr>
<tr>
<td></td>
<td>591</td>
<td>455</td>
<td>1046</td>
</tr>
</tbody>
</table>

- **A** 64
- **B** 455 • \(\frac{64}{136}\)
- **C** 591 • \(\frac{64}{136}\)
- **D** 455 • \(\frac{64}{72}\)
- **E** cannot be determined from the information given

AP2. In a pre-election poll, 13,660 potential voters were categorized based on their annual income (into one of eight categories) and their intended vote (two categories, Republican or Democratic candidate). The resulting chi-square test of independence of these variables gave a test statistic of 270. How many degrees of freedom are there?

- **A** 2
- **B** 7
- **C** 16
- **D** 269
- **E** 13659

AP3. Researchers interested in determining whether there is a relationship between a mother's birthday and the birthday of her oldest child took a random sample of 200 mothers. The mother's birthday was categorized as within a week, more than a week but less than a month, or more than a month from her oldest child's birthday. Assuming that there are 365 days in a year and the same number of people are born on each day, what is the expected frequency of mothers in the “within at most one week” category?

- **A** 8
- **B** 8.219
- **C** 66
- **D** 66.67
- **E** 67

AP4. In a study of whether two acne medications, are equally effective, researchers got eight volunteers and randomly chose one side of each person's face to receive each medication. After one month of use, they counted the number of pimples on each side of each person's face. Which test is most appropriate in this situation?

- **A** matched-pairs t-test
- **B** two-sample t-test
- **C** chi-square test for goodness of fit
- **D** chi-square test for homogeneity of proportions
- **E** chi-square test of independence

AP5. A random sample of 1000 adults from each of the 50 states is taken, and the people are categorized as to whether they have graduated from college or not, generating a 50-by-2 table of counts. Which is the most appropriate test for determining whether the graduation rates differ among the states?

- **A** Compare each count to the expected frequency of 500 and use a chi-square goodness of fit test.
- **B** Use a chi-square test of homogeneity of proportions.
- **C** Use a chi-square test of independence.
- **D** Use a one-proportion z-test.
- **E** Use a t-test for the difference of means.

AP6. For a project, a student plans to roll a six-sided die 1000 times. The teacher becomes suspicious when the student reports getting 168 ones, 165 twos, 170 threes, 167 fours, 164 fives, and 166 sixes. The teacher performs a chi-square test with the alternative hypothesis that the student's reported observed counts are closer to the expected frequencies than can reasonably be attributed to chance. What is the test statistic and P-value for this test?

- **A** \(\chi^2 = 0.14; P-value \approx 0.9996\)
- **B** \(\chi^2 = 0.14; P-value \approx 0.0004\)
- **C** \(\chi^2 = 1000; P-value \approx 0\)
- **D** \(\chi^2 = 23.33; P-value \approx 0.0004\)
- **E** \(\chi^2 = 23.33; P-value \approx 0.9996\)

AP7. A researcher wanted to determine whether Barbarians and Vandals have similar food
preferences. Independent random samples were taken from each tribe and each person was categorized as to whether he or she prefers to eat gruel, turnips, or raw meat. The value of $\chi^2$ was statistically significant. What conclusion can be drawn?

A. The proportion of all Barbarians who prefer gruel is different from the proportion of all Vandals who prefer gruel.

B. The proportion of all Barbarians who prefer gruel is equal to the proportion of all Vandals who prefer gruel.

C. The population proportions are different for at least one of the three choices.

D. The population proportions are different for all three of the choices.

E. The population proportions are equal for all three of the choices.

AP8. Random samples are taken from two populations (Barbarians and Vandals) and each person is categorized as a pillager or a burner. Which of the following, when done appropriately, is equivalent to a chi-square test of homogeneity for this situation?

A. one-sample $z$-test

B. two-sample $z$-test

C. one-sample $t$-test

D. two-sample $t$-test

E. chi-square goodness-of-fit test

Investigative Tasks

AP9. To study the effectiveness of seat belts and air bags, researchers took a careful look at data that were collected on accidents between 1997 and 2002. The data came from a division of the National Highway Traffic Safety Administration that collects detailed data on a random sample of accidents “in which there is a harmful event and from which at least one vehicle is towed.” A total of 22,804 front-seat occupants were involved in the accidents studied. [Source: Mary C. Meyer and Tremika Finney, “Who Wants Airbags?” Chance 18, no. 2 (2005): 3–16. The data here are approximations created from scaling down the population projections in the article.]

a. Do the data in the table show a statistically significant association between seat belt use and surviving an accident?

<table>
<thead>
<tr>
<th></th>
<th>Seat Belts Used</th>
<th>Seat Belts Not Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Killed</td>
<td>35</td>
<td>67</td>
</tr>
<tr>
<td>Survived</td>
<td>16,694</td>
<td>6008</td>
</tr>
</tbody>
</table>

b. Do the data in the table show a statistically significant association between air bag use and surviving an accident?

<table>
<thead>
<tr>
<th></th>
<th>Air Bags Used</th>
<th>Air Bags Not Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Killed</td>
<td>42</td>
<td>60</td>
</tr>
<tr>
<td>Survived</td>
<td>12,315</td>
<td>10,387</td>
</tr>
</tbody>
</table>

c. From the two analyses in parts a and b, which safety feature appears to have the stronger association with survival?

AP10. The analysis in AP9 does not account for seat belts and air bags interacting with each other. This table splits the air bag data according to whether seat belts were also in use during the accident.

<table>
<thead>
<tr>
<th></th>
<th>Seat Belts Used</th>
<th>Seat Belts Not Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Air Bags</td>
<td>19</td>
<td>16</td>
</tr>
<tr>
<td>No Air Bags</td>
<td>10,464</td>
<td>6,230</td>
</tr>
<tr>
<td>Air Bags</td>
<td>23</td>
<td>44</td>
</tr>
<tr>
<td>No Air Bags</td>
<td>1,851</td>
<td>4,157</td>
</tr>
<tr>
<td>Total</td>
<td>10,483</td>
<td>6,246</td>
</tr>
<tr>
<td></td>
<td>1,874</td>
<td>4,201</td>
</tr>
</tbody>
</table>

a. What do the data suggest as to the effect of air bag use when seat belts are also used? Is there a significant association?

b. What do the data suggest about the effect of air bag use when seat belts are not used? Is there a significant association?
Do Mars rocks with more sulfur tend to be redder? Data from the Pathfinder mission to the red planet were used to explore questions like this.
Trivia Question 1: Who are Shark, Barnacle Bill, Half Dome, Wedge, and Yogi?
Answer: Mars rocks, found at the landing site of Mars Pathfinder in July 1997.

Trivia Question 2: Mars is often called the red planet. What makes it red?
Answer: Sulfur?

The Sojourner robot rover rolls out of Mars Pathfinder to meet Barnacle Bill and Yogi.

Display 11.1 shows how the redness of the five Mars rocks is related to their sulfur content. The response variable, redness, is the ratio of red to blue in a spectral analysis; the higher the value, the redder the rock. The explanatory variable is the percentage, by weight, of sulfate in the rock.

Shark: Have you figured out how change in the sulfate content is reflected in change in redness?

Yogi: According to my calculations, the slope of the regression line is 0.525. Rocks that differ by 1 percentage point in their sulfate content will differ by about 0.5, on average, in their measure of redness. I learned how to do this in Chapter 3.

Shark: That’s the story if you take the numbers at face value.

Yogi: Why shouldn’t I?

Shark: Well, after all, we aren’t the only rocks on the planet. The slope, 0.525, is an estimate based on a sample—us! If that cute little rolling Sojourner robot had nuzzled up to different rocks, you would have a different estimate. So the goal of this chapter is to show how to construct confidence intervals and test hypotheses for the slope of a regression line.

Yogi: You need a whole chapter just for that? I know from Chapter 3 how to estimate the slope. So, if you tell me how to find the standard error for the slope, I can do the rest. To get a confidence interval, I just do the usual: $0.525 \pm t^* \cdot SE$. To test a null hypothesis, I simply compute

$$t = \frac{0.525 - \text{hypothesized slope}}{SE}$$

and use the $t$-table to find the $P$-value.

Shark: It sounds familiar, all right. Maybe this can be a short chapter!

This chapter addresses how to make inferences about the unknown true relationship between two quantitative variables. The methods and logic you will learn in this chapter apply to a broad range of such questions.

**In this chapter, you will learn**

- that the slope of a regression line fitted from sample data will vary from sample to sample, and what things affect this variability
- how to estimate the standard error of the slope
- how to construct and interpret a confidence interval for the slope
- how to test whether the slope is different from a hypothesized value
- how to know when to trust confidence intervals and tests
- how to transform variables to make inferences more trustworthy
Variation in the Slope from Sample to Sample

As you know by now, statistical thinking, though powerful, is never as easy or automatic as simply plugging numbers into formulas. In order to use statistical methods appropriately, you need to understand their logic, not just the computing rules. The logic in this chapter is designed for bivariate populations that you can think of as modeled by “linear fit plus random deviation.” That is, all points don’t lie exactly on the line of best fit but cluster around it, forming an elliptical cloud.

Linear Models

In Chapter 3, you learned to summarize a linear relationship with a least squares regression line,

\[ \hat{y} = b_0 + b_1 x \]

That equation is a complete description if you have the entire population, but if you have only a random sample, the values of \( b_0 \) and \( b_1 \) are estimates of the true population parameters. That is, there is some underlying “true” linear relationship that you are trying to estimate, just as you use \( \bar{x} \) as an estimate of \( \mu \). The notation for a linear relationship is

\[ y = (\beta_0 + \beta_1 x) + \varepsilon \]

where \( \beta_0 \) and \( \beta_1 \) refer, respectively, to the intercept and slope of a line that you don’t ordinarily get to see—the true regression line you would get if you had data for the whole population instead of only a sample. The letter \( \varepsilon \) indicates the size of the random deviation—how far a point falls above or below the true regression line. The true regression line, sometimes called the line of means or the line of averages, is written

\[ \mu_y = \beta_0 + \beta_1 x \]

Because such linear models are often used to predict unknown values of \( y \) from known values of \( x \), or to explain how \( x \) influences the variation in \( y \), \( y \) is called the response variable and \( x \) is called the predictor variable or the explanatory variable.

Activity 11.1a is designed to help you understand the roles of these equations and the relationship between them.
ACTIVITY 11.1a  How Fast Do Kids Grow?

What you’ll need:  a copy of Display 11.2

On average, kids from the ages of 8 to 13 grow taller at the rate of 2 in. per year. Heights of 8-year-olds average about 51 in. At each age, the heights are approximately normal, with a standard deviation of roughly 2.1 in. [Source: National Health and Nutrition Examination Survey, May 30, 2000, www.cdc.gov.]

1. Use the information above to fill in the second column of your copy of the table in Display 11.2.

<table>
<thead>
<tr>
<th>Age, x</th>
<th>Average Height from Model, μ_y</th>
<th>Random Deviation, ε</th>
<th>Observed Height, y, of Your Child</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>51</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>13</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Display 11.2  Partial data table for Activity 11.1a.

2. Find the equation of the true regression line, \( \mu_y = \beta_0 + \beta_1 x \), that relates average height, \( y \), to age, \( x \). Interpret \( \beta_0 \) and \( \beta_1 \).

3. Now suppose you have a randomly selected child of each age. To find how much the height of your 8-year-old child deviates from the average height, use your calculator to randomly select a deviation, \( \varepsilon \), from a normal distribution with mean 0 and SD 2.1. [See Calculator Note 11A.] Record \( \varepsilon \) and then add \( \varepsilon \) to the average height. Record the sum in the column labeled “Observed Height.” Repeat, with a new deviation, for each age.

4. Fit a regression equation to the pairs \((x, y)\) where \( y \) is the observed height. Record your estimated slope, \( b_1 \). Is it close to \( \beta_1 \)?

5. Collect the values of \( b_1 \) found by the members of your class and plot these values. Discuss the shape, center, and spread of this plot of estimated slopes.

6. The population of the deviations in step 3 had mean 0. Why?

7. Suppose that, instead of creating your data by simulation, you used actual data by choosing one child from each age group at random and measuring their heights. In what ways is the model in step 3 a reasonable model for this situation? In what ways is it not so reasonable?

**Yogi:** You don’t expect me to believe that real data get created the way we did it in the activity, do you?

**Shark:** No. At best, this model is a good approximation, and it’s up to you to judge how well it describes the process that created your data. But precisely because the model is a simplified version of reality, when it’s reasonable, it’s quite useful.
The conditional distribution of $y$ given $x$ refers to all the values of $y$ for a fixed value of $x$. For example, think of the distribution of children's heights described in Activity 11.1a. If you were to make a scatterplot of (age, height) for a sample of hundreds of children, the vertical column of points for all the 8-year-old children would approximate the conditional distribution of the height, $y$, given that the age, $x$, is 8, as shown in Display 11.3. Each conditional distribution of height for a given age has a mean, called $\mu_y$ for a population and $\bar{y}$ for a sample, and a measure of variability, called $\sigma$ for a population and $s$ for a sample. A linear model is appropriate for a set of data if the conditional means fall near a line and the variability is about the same for each conditional distribution.

Display 11.3  Scatterplot of height versus age. Conditional distributions of children's height given their age are the vertical columns of dots.

**Linear Models**

D1. Refer to the information about children's heights in Activity 11.1a. What are the mean and standard deviation of the conditional distribution of the heights given that the age is 10? Given that the age is 12?

D2. Explain the difference between the linear model $y = \beta_0 + \beta_1 x + \epsilon$ and the fitted equation $\hat{y} = b_0 + b_1 x$. In particular, what is the difference between $\beta_1$ and $b_1$? What is the difference between a random deviation $\epsilon$ from the model and an observed residual, $y - \hat{y}$?

**The Variability of $b_1$ from Sample to Sample**

In Activity 11.1a, you actually knew the value of the theoretical slope, $\beta_1$, so when you found an estimate $b_1$ based on the data from your sample, you could tell how far that estimate was from the true value. In real situations, of course, you don’t know the true slope. You must use your data to estimate not only $\beta_1$ but also the standard error of $b_1$. Then, if a linear model is appropriate, you can rely on the fact that 95% of the time the estimated slope, $b_1$, will be within approximately two standard errors of the true slope, $\beta_1$.

To move toward finding a formula for the standard error of $b_1$, think about the situations when $b_1$ would tend to be close to $\beta_1$ and when it would tend to be farther away. As you might have guessed, with a larger sample size, $b_1$ tends to be closer to $\beta_1$. Activity 11.1b is designed to help you see what other factors affect the variation in $b_1$ when the standard model (true regression line plus random variation) applies.
What Affects the Variation in $b_1$?

In this activity, you will create four scatterplots. For each plot, there will be only two values of the predictor, $x$, and you will generate four values of the response, $y$, for each $x$. These response values will come from normal distributions with the specified means and standard deviations.

1. For each value of $x$ given in Cases 1–4, generate four values of $y$ from a normal distribution with the given mean and standard deviation. [See Calculator Note 11A to review how to generate random numbers from a normal distribution, and see Calculator Note 11B to learn how to augment a list, which will speed up your work on this step.] Plot the resulting eight ordered pairs for each case on a scatterplot. Fit a regression line and record the slope.

   **Case 1**
   - $x = 0$; conditional distribution of $y$ has $\mu_y = 10$ and $\sigma = 3$
   - $x = 1$; conditional distribution of $y$ has $\mu_y = 12$ and $\sigma = 3$

   **Case 2**
   - $x = 0$; conditional distribution of $y$ has $\mu_y = 10$ and $\sigma = 3$
   - $x = 4$; conditional distribution of $y$ has $\mu_y = 18$ and $\sigma = 3$

   **Case 3**
   - $x = 0$; conditional distribution of $y$ has $\mu_y = 10$ and $\sigma = 5$
   - $x = 1$; conditional distribution of $y$ has $\mu_y = 12$ and $\sigma = 5$

   **Case 4**
   - $x = 0$; conditional distribution of $y$ has $\mu_y = 10$ and $\sigma = 5$
   - $x = 4$; conditional distribution of $y$ has $\mu_y = 18$ and $\sigma = 5$

2. Collect the estimated slopes from your class. Construct four dot plots of these estimated slopes, one plot for each case. What is the theoretical slope, $\beta_1$, in each case? For each case, estimate the mean of the distribution of estimated slopes and compare it to $\beta_1$.

3. How does the variation in the estimated slopes change with an increase in the variation of the response variable, $y$? How does the variation in the estimated slopes change with an increase in the spread of the explanatory variable, $x$?

You learned in Activity 11.1b that the variation in the estimated slopes depends not only on how much the values of $y$ vary for each fixed value of $x$ but also on the spread of the $x$-values.

In the methods of inference in this chapter, the variability in $y$ is assumed to be the same for each conditional distribution. That is, if you picked a value of $x$ and computed the standard deviation of all the associated values of $y$ in the population, you’d get the same number, $\sigma$, as you would if you picked any other $x$, as shown in Display 11.4. This implies that $\sigma$ also measures the variability of all values of $y$ about the true regression line. You can use this fact to estimate $\sigma$ from your data.
Variability in $x$ and $y$

The common variability of $y$ at each $x$ is called $\sigma$. It is estimated by $s$, which can be thought of as the standard deviation of the residuals. You compute $s$ using all $n$ values of $y$ in the sample:

$$\sigma \approx s = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - 2}} = \sqrt{\frac{\text{SSE}}{n - 2}}$$

The spread in the values of $x$ is measured by $\sqrt{\sum (x_i - \bar{x})^2}$, where $\bar{x}$ is computed using all $n$ values of $x$ in the sample.

Display 11.5, which shows some results from Activity 11.1b, shows that the variability in the slope depends on the sample size and the variability in $x$ and $y$.

Plot I: $x = 0, \mu_y = 10; x = 1, \mu_y = 12; \sigma = 3$

Plot II: $x = 0, \mu_y = 10; x = 4, \mu_y = 18; \sigma = 3$

Plot III: $x = 0, \mu_y = 10; x = 1, \mu_y = 12; \sigma = 5$

Plot IV: $x = 0, \mu_y = 10; x = 4, \mu_y = 18; \sigma = 5$

Display 11.5 Regression lines relating variability in $b_1$ to variation in $y$ and to spread in $x$. 

Note that $s$ measures the variability of the residuals.
Each plot in Display 11.5 shows five regression lines for one of the four cases; for example, Plot I shows five regression lines for Case 1. Each line was constructed using the instructions in step 1 of the activity. Plots I and II (or III and IV) show that a wider spread in \( x \) results in regression lines with less variability in their slope—even though the values of \( y \) vary equally for each \( x \)-value. By comparing plot I with plot III (or plot II with plot IV), you can see something more expected: More variability in each conditional distribution of \( y \) means more variability in the slope, \( b_1 \).

**DISCUSSION**

The Variability of \( b_1 \) from Sample to Sample

D3. In the “scatterplots” in Display 11.6, imagine taking a sample of responses within each rectangle. That is, the rectangles define the regions in which the responses lie for each value of \( x \). Suppose you take a sample and calculate a regression line. Then you repeat the process many times. Which of the three “plots” should produce regression lines with the smallest variation in slope? Which should produce regression lines with the largest variation in slope?

![Display 11.6 Three “scatterplots” for D3.](image)

D4. If samples were to be selected from the rectangles in Display 11.7, would you have any concerns about using a straight line to model the relationship between \( x \) and \( y \)? Why or why not?

![Display 11.7 A “scatterplot” for D4.](image)
The Standard Error of the Slope

You now know what affects how much the slope, \( b_1 \), of the regression line varies from sample to sample: the sample size, \( n \); the variability, \( \sigma \), of \( y \) at each fixed value of \( x \) (estimated by \( s \)); and the spread in the values of \( x \). This box gives the formula for estimating the standard error, \( s_{b_1} \), of the sampling distribution of the slope.

**Formula for Estimating the Standard Error of the Slope**

\[
s_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{\sqrt{\sum (y_i - \hat{y}_i)^2}}{n - 2} \frac{\sqrt{\sum (x_i - \bar{x})^2}}{\sum (x_i - \bar{x})^2}
\]

The formula does what you would expect: The slope varies less from sample to sample when the sample size is larger, when the values of \( y \) tend to be closer to the regression line, and when the values of \( x \) are more spread out.

**Example: French Fries**

A statistics class at the University of Wisconsin–Stevens Point decided to study the mass and number of fries in small and large bags of french fries at McDonald’s. They bought 32 bags of each size during two different time periods on two consecutive days at the same McDonald’s and weighed the fries. The data are given in Display 11.8. (Note that data are missing for three of the large bags due to various problems with the data collection.)

Study the issue of predicting the total mass of the bag of fries from the number of fries by answering a–c.

a. Plot the data and, if appropriate, fit least squares regression lines for each bag size. Interpret the slope for the small bags and then compare with the slope for the large bags.

b. Study the residual plots. Do you see any pattern that suggests that a “line plus random error” is a poor model?

c. Calculate the standard error of the slope for the small bags and interpret it. Then calculate the standard error of the slope for the large bags and compare with that for the small bags.
Display 11.8  Number and mass of french fries. [Source: Nathan Wetzel, “McDonald’s French Fries. Would You Like Small or Large Fries?” STATS, 43 (Spring 2005): 12–14.]
Solution

a. The plots, shown in Display 11.9, indicate a positive linear trend, even though there is considerable variation around the regression lines. The slope for the small bags means that if one bag has 1 more fry than another, you would expect the bag of fries to weigh 0.741 gram more. In other words, a fry weighs about 0.741 gram. The slope for the large bags is 0.321, so the estimate of the weight of a fry is less than half that from the small bags.

\[
\text{Small Bag Mass} = 0.741 \times \text{Small Bag Number} + 38; r^2 = 0.40
\]

\[
\text{Large Bag Mass} = 0.321 \times \text{Large Bag Number} + 115; r^2 = 0.23
\]

Display 11.9 Regression of total mass versus number of fries in bag.

b. The residual plots look much like randomly scattered points, although the small fries have some potentially influential observations because of the large residuals, while the large fries have a potentially influential observation because of a large gap on the x-axis.
c. The slope of the least squares regression line for small bags has an estimated standard error given by

\[
s_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{\sqrt{\Sigma (y_i - \hat{y}_i)^2}}{n - 2} = \frac{\sqrt{2057.3}}{32 - 2} = \frac{\sqrt{2499.5}}{} = 0.1656
\]

So, if you were to take many random samples of 32 bags of small fries and compute the slope of the regression line for predicting mass from number of fries, the standard deviation of the distribution of these slopes would be about 0.1656.

For the large bags, the estimated standard error is

\[
s_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{\sqrt{\Sigma (y_i - \hat{y}_i)^2}}{n - 2} = \frac{\sqrt{3337.10}}{29 - 2} = \frac{\sqrt{9886.69}}{} = 0.1118
\]

The estimated standard error of the slope for the large bags is smaller than that for the small bags, even though there is much more variation in the masses \((y)\) of the large bags. This happens because the larger variation in masses is offset by the larger variation in the numbers of fries \((x)\) for the large bags.

**DISCUSSION**

**The Standard Error of the Slope**

D5. You plan to collect data in order to predict the actual temperature of your oven from the temperature shown on its thermostat. Suppose this relationship is linear. How will you design this study? What have you learned in this section that will help you?

D6. Use Display 11.5 on page 743 to explain why it makes sense for the expression with the \(x\)'s to be in the denominator of the formula for \(s_{b_1}\).

**Summary 11.1: Variation in the Slope from Sample to Sample**

Suppose you want to estimate an underlying linear relationship, \(y = \beta_0 + \beta_1 x + \varepsilon\) where \(\mu_y = \beta_0 + \beta_1 x\) is the equation of the true regression line. The intercept, \(b_0\), and slope, \(b_1\), in the least squares regression equation \(\hat{y} = b_0 + b_1 x\) serve as your estimates of the population parameters \(\beta_0\) and \(\beta_1\). Typically, you will be most interested in how closely \(b_1\) estimates \(\beta_1\). Use the values in your sample and this formula to estimate the standard error of the slope:

\[
s_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} = \sqrt{\frac{\Sigma (y_i - \hat{y}_i)^2}{n - 2}} = \sqrt{\frac{\Sigma (y_i - \hat{y}_i)^2}{\Sigma (x_i - \bar{x})^2}}
\]
The slope, $b_1$, of the regression line varies less from sample to sample when

- the sample size is larger
- the residuals are smaller
- the values of $x$ are farther apart

In the following sections, you will learn to compute confidence intervals and test statistics for the slope, but your conclusions won't be valid unless "line plus random deviations" is the appropriate model for your data. So always plot your data first to judge whether a line is a reasonable model.

### Practice

#### Linear Models

**P1.** The scatterplot of data on pizzas in Display 3.31 on page 134 shows the number of calories versus the number of grams of fat in one serving of several kinds of pizza. Fat contains 9 calories per gram.

a. What would be the theoretical slope of a line representing such data?

b. What does the intercept of the line tell you?

c. What are some reasons why not all of the points fall exactly on a line with the slope in part a?

**P2.** According to Leonardo da Vinci, a person's arm span and height are about equal. Display 11.10 gives height and arm span measurements for a sample of 15 high school students.

<table>
<thead>
<tr>
<th>Arm Span (cm)</th>
<th>Height (cm)</th>
<th>Arm Span (cm)</th>
<th>Height (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>168.0</td>
<td>170.5</td>
<td>129.0</td>
<td>132.5</td>
</tr>
<tr>
<td>172.0</td>
<td>170.0</td>
<td>169.0</td>
<td>165.0</td>
</tr>
<tr>
<td>101.0</td>
<td>107.0</td>
<td>175.0</td>
<td>179.0</td>
</tr>
<tr>
<td>161.0</td>
<td>159.0</td>
<td>154.0</td>
<td>149.0</td>
</tr>
<tr>
<td>166.0</td>
<td>166.0</td>
<td>142.0</td>
<td>143.0</td>
</tr>
<tr>
<td>174.0</td>
<td>175.0</td>
<td>156.5</td>
<td>158.0</td>
</tr>
<tr>
<td>153.5</td>
<td>158.0</td>
<td>164.0</td>
<td>161.0</td>
</tr>
<tr>
<td>95.0</td>
<td>95.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Display 11.10** Height versus arm span for 15 high school students.

a. What is the theoretical regression line, $\mu_y = \beta_0 + \beta_1 x$, that Leonardo is proposing for this situation? Use arm span as the explanatory variable.

b. Find the least squares regression line, $\hat{y} = b_0 + b_1 x$, for these data. Interpret the slope. Compare the estimated slope and intercept to the theoretical slope and intercept in part a. Are they close?

c. Calculate the random deviation, $e$, for each student, using the theoretical regression line. Plot these random deviations against the arm spans and comment on the pattern.

d. For each student, calculate the residual from the estimated regression line. Plot the residuals against the arm spans and comment on the pattern. Is the pattern similar to that for the random deviations?

**P3.** Every spring, visitors eagerly await the opening of the spectacular Going-to-the-Sun Road in Glacier National Park, Montana. A typical range of yearly snowfall for the area is from 30 to 70 inches. The amount of snow is measured at Flattop Mountain, near the top of the road, on the first Monday in April and is given in swe (snow water equivalent: the water content obtained from melting). From analysis of past data, when the amount of snow was 30 inches of swe, the road opened, on average, on the 150th day of the year. Every additional 0.57 in. of swe measured at Flattop Mountain meant another day on average until the road opened. [Source: “Spring Opening of the Going-to-the-Sun Road and Flattop Mountain SNOTEL Data,” Northern Rocky Mountain Science Center, U.S. Geological Survey, December 2001, www.nrmsc.usgs.gov.]

a. If you write an equation summarizing the given information, should you use...
the form \( y = \beta_0 + \beta_1 x \) or \( \hat{y} = b_0 + b_1 x \)? Explain.

b. Write an equation that predicts when the road will be open, given the swe. In this situation, what is the response variable, \( y \)? The predictor variable, \( x \)?

c. In 2005, the Flattop Mountain station recorded 31.0 inches of swe. What date would you predict the road opened?

d. Do you think the random deviation, \( \varepsilon \), in this situation tends to be relatively large or small? Make a guess as to what it might be on average.

The Variability of \( b_1 \) from Sample to Sample

P4. Refer to the table in Display 11.1 on page 737, which gives the sulfate percentage, \( x \), and redness, \( y \), of the Mars rocks. Compute \( s \), the estimate of the common variability of \( y \) at each \( x \).

P5. Five quantities are listed here, along with two possible values of each. For each quantity, decide which value will give you the larger variability in \( b_1 \) (assuming all other things stay the same), and give a reason why.
   a. the standard deviation, \( \sigma \), of the individual response values of \( y \) at each value of \( x \): 3 or 5
   b. the spread of the \( x \)-values: 3 or 10
   c. the number of observations, \( n \): 10 or 20
   d. the true slope, \( \beta_1 \): 1 or 3
   e. the true intercept, \( \beta_0 \): 1 or 7

The Standard Error of the Slope

P6. Look again at the Mars rocks data in Display 11.1 on page 737.
   a. Compute the estimate of the standard error of the slope, \( \sigma_{b_1} \).
   b. On the computer output for the regression in Display 11.11, locate the standard error of the slope. Then locate your estimate from P4 of the variation in \( y \) about the line. What is the equation of the regression line?

Display 11.11 Regression analysis for the Mars rocks data.

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1.48269</td>
<td>1</td>
<td>1.48269</td>
<td>12.7</td>
</tr>
<tr>
<td>Residual</td>
<td>0.349593</td>
<td>3</td>
<td>0.116531</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>s.e. of Coeff</th>
<th>t-ratio</th>
<th>prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1.71525</td>
<td>0.4010</td>
<td>4.28</td>
<td>0.0235</td>
</tr>
<tr>
<td>Sulfate</td>
<td>0.524901</td>
<td>0.1472</td>
<td>3.57</td>
<td>0.0576</td>
</tr>
</tbody>
</table>

Display 11.12 Data and scatterplot for P7.

a. Find both the true regression line, \( \mu_y = \beta_0 + \beta_1 x \), and the least squares regression line, \( \hat{y} = b_0 + b_1 x \), and compare them.
b. Compute $s$ and compare it to $\sigma = 3$.

c. Compute $s_{b_1}$, the estimate of the standard error of $b_1$, and compare it to the true value of the standard error. (Note: The true value of the SE for $b_1$ is $\frac{\sigma}{\sqrt{\sum(x - \bar{x})^2}}$).

d. Does the least squares regression line go through the mean of the responses at both $x = 0$ and $x = 4$?

P8. Each of the lists in I–V gives the values of $x$ used to compute a regression line. Assuming $\sigma = 1$ in all cases, order the lists from the one with the largest standard error of the slope to the one with the smallest. (You should be able to rank these without doing much, if any, computation.)

<table>
<thead>
<tr>
<th>List</th>
<th>$x$</th>
<th>$\sigma$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I.</td>
<td>1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>II.</td>
<td>1, 2, 3, 4, 5, 6, 7, 8, 9, 10</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>III.</td>
<td>1, 1, 1, 1, 1, 12, 12, 12, 12, 12</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>IV.</td>
<td>1, 1, 4, 4, 4, 9, 9, 9, 12, 12, 12</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>V.</td>
<td>1, 2, 3, 4, 5, 6, 7</td>
<td>1</td>
<td>7</td>
</tr>
</tbody>
</table>

P9. As you did in P8, order these lists from the largest to the smallest value of the true $SE$ of the slope. (For these, you might need to use the formula and do some computation.)

<table>
<thead>
<tr>
<th>List</th>
<th>$x$</th>
<th>$\sigma$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I.</td>
<td>2; 1, 1, 1, 1, 3, 3, 3</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>II.</td>
<td>1; 1, 1, 1, 2, 2, 2, 2, 3, 3, 3</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>III.</td>
<td>3; 1, 1, 2, 2, 3, 4, 4, 5, 5, 6, 6</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>IV.</td>
<td>1; 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6</td>
<td>1</td>
<td>11</td>
</tr>
</tbody>
</table>

Exercises

E1. Display 11.13 gives some information about a random sample of motor vehicle models commonly sold in the United States.

<table>
<thead>
<tr>
<th>Manufacturer</th>
<th>Model</th>
<th>price</th>
<th>mpg</th>
<th>liter</th>
<th>hp</th>
<th>rpm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chevrolet</td>
<td>Cavalier</td>
<td>13.4</td>
<td>36</td>
<td>2.2</td>
<td>110</td>
<td>5200</td>
</tr>
<tr>
<td>Chevrolet</td>
<td>Lumina APV</td>
<td>16.3</td>
<td>23</td>
<td>3.8</td>
<td>170</td>
<td>4800</td>
</tr>
<tr>
<td>Chevrolet</td>
<td>Astro</td>
<td>16.6</td>
<td>20</td>
<td>4.3</td>
<td>165</td>
<td>4000</td>
</tr>
<tr>
<td>Dodge</td>
<td>Shadow</td>
<td>11.3</td>
<td>29</td>
<td>2.2</td>
<td>93</td>
<td>4800</td>
</tr>
<tr>
<td>Dodge</td>
<td>Caravan</td>
<td>19.0</td>
<td>21</td>
<td>3.0</td>
<td>142</td>
<td>5000</td>
</tr>
<tr>
<td>Eagle</td>
<td>Vision</td>
<td>19.3</td>
<td>28</td>
<td>3.5</td>
<td>214</td>
<td>5800</td>
</tr>
<tr>
<td>Ford</td>
<td>Probe</td>
<td>14.0</td>
<td>30</td>
<td>2.0</td>
<td>115</td>
<td>5500</td>
</tr>
<tr>
<td>Hyundai</td>
<td>Elantra</td>
<td>10.0</td>
<td>29</td>
<td>1.8</td>
<td>124</td>
<td>6000</td>
</tr>
<tr>
<td>Lexus</td>
<td>SC300</td>
<td>35.2</td>
<td>23</td>
<td>3.0</td>
<td>225</td>
<td>6000</td>
</tr>
<tr>
<td>Mazda</td>
<td>RX-7</td>
<td>32.5</td>
<td>25</td>
<td>1.3</td>
<td>255</td>
<td>6500</td>
</tr>
<tr>
<td>Oldsmobile</td>
<td>Achieva</td>
<td>13.5</td>
<td>31</td>
<td>2.3</td>
<td>155</td>
<td>6000</td>
</tr>
<tr>
<td>Pontiac</td>
<td>Grand Prix</td>
<td>18.5</td>
<td>27</td>
<td>3.4</td>
<td>200</td>
<td>5000</td>
</tr>
<tr>
<td>Suzuki</td>
<td>Swift</td>
<td>8.6</td>
<td>43</td>
<td>1.3</td>
<td>70</td>
<td>6000</td>
</tr>
<tr>
<td>Volkswagen</td>
<td>Fox</td>
<td>9.1</td>
<td>33</td>
<td>1.8</td>
<td>81</td>
<td>5500</td>
</tr>
<tr>
<td>Volvo</td>
<td>850</td>
<td>26.7</td>
<td>28</td>
<td>2.4</td>
<td>168</td>
<td>6200</td>
</tr>
</tbody>
</table>

Display 11.13 Table and scatterplots of car models data. [Source: Journal of Statistics Education Data Archives, June 2002.]

The variables are:

- price typical selling price, in thousands of dollars
- mpg typical highway mileage, in miles per gallon
- liter size of the engine, in liters
- hp horsepower rating of the vehicle
- rpm maximum revolutions per minute the engine is designed to produce
Consider these pairs of variables:
\[ x = \text{hp}, \ y = \text{price} \quad x = \text{hp}, \ y = \text{mpg} \]
\[ x = \text{liter}, \ y = \text{mpg} \quad x = \text{rpm}, \ y = \text{mpg} \]

a. Refer to the scatterplot for each pair of variables and tell whether you think a line gives a suitable summary of their relationship.

b. Compute the slope of the regression line for \( x = \text{hp} \) and \( y = \text{mpg} \). Interpret this slope in context. Compute the estimated standard error of the slope.

c. Find your values from part c on the computer printout in Display 11.14. Also find and interpret \( s \), the estimate of \( \sigma \).

Bivariate Fit of Miles per Gallon by Horsepower

![Bivariate Fit of Miles per Gallon by Horsepower](image)

Linear Fit
\[ \text{MPG} = 38.980466 - 0.0693953 \times \text{HP} \]

Summary of Fit
- RSquare: 0.405843
- RSquare Adj: 0.360138
- Root Mean Square Error: 4.778485
- Mean of Response: 28.4
- Observations (or Sum Wgts): 15

Analysis of Variance
- Source: Model, Error, Total
- DF: 1, 13, 14
- Sum of Squares: 202.75910, 296.84090, 499.60000
- Mean Square: 202.759, 22.834, 0.0106
- F Ratio: 8.8797
- Prob>F: 0.0106

Parameter Estimates
- Intercept: 38.980466, Std Error: 3.758830, t Ratio: 10.37, Prob>|t|: <.0001
- HP: -0.0693953, Std Error: 0.023288, t Ratio: -2.98, Prob>|t|: 0.00106

Display 11.14 Regression analysis for the car models problem.

E2. How does energy consumption in a residence relate to the outdoor temperature? Some interesting patterns can be detected through study of available monthly records on natural gas and electricity usage for a gas-heated single-family residence (with no air conditioning) in the Boston area, along with monthly data on outdoor temperature and heating degree days. Heating degrees for a day with mean temperature less than 65°F are defined as \((65 - \text{mean daily temperature})\). The data are given in Display 11.15, with these variables:

- **Mean Temp**: mean monthly temperature in Boston, in degrees Fahrenheit
- **Mean Gas**: mean natural gas usage per day for the month, in therms
- **Mean KWH**: mean electricity usage per day for the month, in kilowatt-hours
- **Heat DD**: total heating degree days for the month

<table>
<thead>
<tr>
<th>Month</th>
<th>Mean Temp</th>
<th>Mean Gas</th>
<th>Mean KWH</th>
<th>Heat DD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jun-95</td>
<td>69</td>
<td>1.3</td>
<td>11.9</td>
<td>30</td>
</tr>
<tr>
<td>Aug-95</td>
<td>73</td>
<td>0.7</td>
<td>15.6</td>
<td>2</td>
</tr>
<tr>
<td>Oct-95</td>
<td>59</td>
<td>1.5</td>
<td>18</td>
<td>214</td>
</tr>
<tr>
<td>Nov-95</td>
<td>46</td>
<td>4.5</td>
<td>14.9</td>
<td>685</td>
</tr>
<tr>
<td>Dec-95</td>
<td>29</td>
<td>8.9</td>
<td>18.1</td>
<td>1023</td>
</tr>
<tr>
<td>Jan-96</td>
<td>30</td>
<td>11.6</td>
<td>18.1</td>
<td>1074</td>
</tr>
<tr>
<td>Feb-96</td>
<td>31</td>
<td>10.7</td>
<td>18.8</td>
<td>981</td>
</tr>
<tr>
<td>Mar-96</td>
<td>37</td>
<td>11.6</td>
<td>37.8</td>
<td>875</td>
</tr>
<tr>
<td>Apr-96</td>
<td>48</td>
<td>7.5</td>
<td>17.6</td>
<td>510</td>
</tr>
<tr>
<td>May-96</td>
<td>57</td>
<td>3.5</td>
<td>17.9</td>
<td>264</td>
</tr>
<tr>
<td>Jun-96</td>
<td>68</td>
<td>1.5</td>
<td>13</td>
<td>20</td>
</tr>
<tr>
<td>Jul-96</td>
<td>71</td>
<td>0.8</td>
<td>17.4</td>
<td>6</td>
</tr>
<tr>
<td>Aug-96</td>
<td>53</td>
<td>1.9</td>
<td>17.7</td>
<td>358</td>
</tr>
<tr>
<td>Sep-96</td>
<td>40</td>
<td>5</td>
<td>22.3</td>
<td>739</td>
</tr>
<tr>
<td>Oct-96</td>
<td>39</td>
<td>7.3</td>
<td>20.7</td>
<td>792</td>
</tr>
<tr>
<td>Nov-96</td>
<td>29</td>
<td>9.3</td>
<td>32.8</td>
<td>1104</td>
</tr>
<tr>
<td>Dec-96</td>
<td>36</td>
<td>9.7</td>
<td>30.5</td>
<td>806</td>
</tr>
<tr>
<td>Jan-97</td>
<td>37</td>
<td>7.9</td>
<td>24.6</td>
<td>868</td>
</tr>
<tr>
<td>Feb-97</td>
<td>46</td>
<td>5.8</td>
<td>17.2</td>
<td>551</td>
</tr>
<tr>
<td>Mar-97</td>
<td>56</td>
<td>3.2</td>
<td>26.2</td>
<td>269</td>
</tr>
</tbody>
</table>

(continued)
a. Refer to the scatterplots in Display 11.15. For each pair of variables, tell whether you think a line gives a suitable summary of the relationship.

b. By looking at the scatterplots, estimate which of the four pairs of variables has the largest standard error of the slope and which has the smallest.

c. Compute the slope of the least squares regression line for the relationship between \( y = Mean\ Gas \) and \( x = Mean\ Temp \). Interpret the slope in context. Compute the estimated standard error of the slope.

d. Find your values from part c on the computer printout in Display 11.16. Also find and interpret \( s \), the estimate of \( \sigma \).

a. What is the theoretical slope of a line fit to such data? Interpret this slope in context.

b. What are some reasons why not all of the points fall exactly on a line with the slope in part a?

E4. Suppose you have a collection of equilateral triangles of different sizes, as in the structural steel of a bridge support. You have both the measured areas and the measured side lengths of these triangles, and you plot the area inside the triangle against the square of the length of a side.

E5. Display 11.17 gives the sulfate content and redness of six soil samples from Mars.

a. Compare the plot in Display 11.17 to the one for the Mars rocks in Display 11.1 on page 737. Which value of $s_b$, do you expect to be larger—the one for the five rocks, or the one for the six soil samples? Why?

b. Compute $s_b$ for the soil samples, and check your conjecture from part a. (You computed $s_b$ for the rocks in P6.)

E6. In the Mars soil sample data of Display 11.17, one of the points appears to be an outlier. Suppose that point is removed.

a. Suppose a line is fit to the remaining five points. Do you expect the estimated SE of the slope to be larger for the original data or for the revised data?

b. Compute $s_b$ using the remaining five points, and check your conjecture from part a. (You computed $s_b$ for all size points in E5.)

E7. In Activity 11.1a about the heights of children ages 8 to 13, the response variable is height, the predictor variable is age, and the average height of 8-year-olds is 51 in. Begin by assuming that, on average children grow 2 in. per year and that, at any one age, the distribution of heights is approximately normal. Sketch a scatterplot that shows the sort of data you would expect to get in step 3 of Activity 11.1a if you used these standard deviations instead of 2.1.

a. $\sigma = 3$

b. $\sigma = 1$
c. This time, use a different value of \( \sigma \) for different ages: age 8, \( \sigma = 1 \text{ in.} \); age 9, \( \sigma = 2 \text{ in.} \); age 10, \( \sigma = 3 \text{ in.} \); . . . ; age 13, \( \sigma = 6 \text{ in.} \).

E8. Refer to E7. Sketch a scatterplot that shows the sort of data you would expect to get in step 3 of Activity 11.1a if you used these new theoretical models.

a. With \( \sigma = 3 \), assume that children grow faster and faster as they get older so that the average height at age \( x \) is given by \( 50 + (x - 8)^2 \).

b. Keep everything the same as in part a, except this time use the values of \( \sigma \) from E7, part c.

E9. Suppose you were to create a model for the Mars rocks data in Display 11.1 on page 737 using the same approach as in Activity 11.1a. Write out directions telling how to use the model to find the redness of a rock that has the same sulfur content as Half Dome. (To do this, you will need to make some assumptions about things you don't know.) In what ways do you consider your model a reasonable one for the Mars rocks data? In what ways is it not so reasonable?

E10. As in P8, order these lists according to the size of the standard error of the slope. (No computation should be necessary.)

I. \( \sigma = 5; x = 1, 1, 1, 3, 3, 3, 3 \)
II. \( \sigma = 2; x = 1, 1, 1, 2, 2, 2, 3, 3, 3, 3 \)
III. \( \sigma = 3; x = 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6 \)
IV. \( \sigma = 2; x = 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6 \)
V. \( \sigma = 7; x = 1, 2, 2, 3 \)

### 11.2 Making Inferences About Slopes

The first part of this section presents and illustrates a significance test for a slope, and the second part gives a confidence interval. The logic here is important, but you’ve seen it before, and the computations follow a familiar pattern. Pay special attention to checking how well the model fits the data. This is the part of inference that separates humans from computers. Any old box of microchips can compute a \( P \)-value, but it takes experience and judgment to figure out whether the \( P \)-value really tells you anything.

#### The Test Statistic for a Slope

Often there is no positive or negative linear relationship between two variables. This is true, for example, of the sum of the last four digits of a person’s phone number and the number of letters in his or her full name, as you will see in Activity 11.2a.

### Activity 11.2a Phone Numbers and Names

Let \( x \) be the sum of the last four digits of a person’s phone number and \( y \) be the number of letters in that person’s full name. Record \( x \) and \( y \) for each student in your class. Then make a scatterplot and compute the equation of the regression line.
Although the scatterplot from Activity 11.2a probably shows little or no obvious association, chances are excellent that $b_1$ isn’t exactly equal to 0. Even when the true slope, $\beta_1$, is 0, the estimate, $b_1$, will usually turn out to be different from 0. In such cases, the estimated slope is not “significant” and differs from 0 simply because cases were picked at random. If another class did the activity, the value of $b_1$ probably would not be 0 either and probably would be different from the value of $b_1$ for your class as well. A significance test for the slope of a regression line asks, “Is that trend real, or could the numbers come out the way they did by chance?”

The test statistic is based on how far $b_1$ is from 0 (or some other hypothesized value of $\beta_1$) in terms of the standard error.

**The test statistic for the slope** is the difference between the slope, $b_1$, estimated from the sample, and the hypothesized slope, $\beta_{10}$, measured in standard errors:

$$t = \frac{b_1 - \beta_{10}}{s_{b_1}}$$

If a linear model is correct and the null hypothesis is true, then the test statistic has a $t$-distribution with $n - 2$ degrees of freedom.

**Yogi:** What a relief! I was waiting for them to introduce yet another new distribution: the $z$, the $t$, the $\chi^2$, the blah-blah-cube.

**Shark:** But aren’t you worried about one little thing? Why does this statistic have a $t$-distribution and not something else?

**Yogi:** Worried? I said I was relieved! I know where the $t$-distribution is on my calculator. The $df$ rule is easy. I am happy, and now you are trying to raise problems. You have not had enough sulfur in your diet.

**Shark:** Okay, we’ll let it go until your next course in statistics. For now, simply notice that the sample slope behaves something like a mean.

---

**DISCUSSION**

The Test Statistic for a Slope

D7. Using the printout for the Mars rocks data in Display 11.11 on page 750, verify the value of the $t$-statistic. Then verify $df$, and use Table B on page 826 to check that the $P$-value in the computer printout is consistent with that in the table.

D8. Display 11.18 shows the combined data for the five Mars rocks and six soil samples. In relation to the $t$-statistics
for the rocks and for the soil samples, where do you expect the value of the \( t \)-statistic for the combined data to fall? Why?

Display 11.18  *Redness versus sulfate percentage* for the five Mars rocks and six soil samples together.

**Significance Test for a Slope**

The box describes the steps for testing the significance of a slope. Generally, you will use this test when you have bivariate data from a sample that appear to have a positive (or negative) linear association and you want to establish that this association is “real.” That is, you want to determine that the nonzero slope you see didn’t happen just by chance—that there actually is a true linear relationship with a nonzero slope so knowing the value of \( x \) is helpful in predicting the value of \( y \).

**Components of a Significance Test for a Slope**

1. **Check conditions.** For the test to work, the conditional distributions of \( y \) for fixed values of \( x \) must be approximately normal, with means that lie on a line and standard deviations that are constant across all values of \( x \). Of course, you can’t check the population for this; you have to use the sample. You need to do several things to check conditions:
   - Verify that you have one of these situations:
     - a single random sample from a bivariate population
     - a set of independent random samples, one for each fixed value of the explanatory variable, \( x \)
     - an experiment with a random assignment of treatments to units
   - Make a scatterplot and check to see if the relationship looks linear.
   - Make a residual plot to check departures from linearity and that the residuals are of uniform size across all values of \( x \).
   - Make a univariate plot (dot plot, stemplot, or boxplot) of the residuals to see if it’s reasonable to assume that they came from a normal distribution.

[See Calculator Note 11C to learn how your calculator can help you check conditions.]
2. **State the hypotheses.** The null and alternative hypotheses usually will be
   \[ H_0: \beta_1 = 0 \text{ and } H_1: \beta_1 \neq 0, \]
   where \( \beta_1 \) is the slope of the true regression line. However, the test may be one-sided, and the hypothesized value, \( \beta_1 \), may be some constant other than 0.

3. **Compute the value of the test statistic, find the \( P \)-value, and draw a sketch.** The test statistic is
   \[ t = \frac{b_1 - \beta_{1_0}}{s_{b_1}} \]
   Here \( b_1 \) and \( s_{b_1} \) are computed from your sample. To find the \( P \)-value, use your calculator’s \( t \)-distribution with \( n - 2 \) degrees of freedom, where \( n \) is the number of ordered pairs in your sample. [See Calculator Note 11D.]

4. **Write your conclusion linked to your computations and in context.** The smaller the \( P \)-value, the stronger the evidence against the null hypothesis. Reject \( H_0 \) if the \( P \)-value is less than the given value of \( \alpha \), typically 0.05. Alternatively, compare the value of \( t \) to the critical value, \( t^* \). Reject \( H_0 \) if \( |t| \geq t^* \), for a two-sided test.

**Yogi:** *Four conditions! What happened to good old “line plus random variation”?!*

**Shark:** It’s still there. But “variation” takes in a lot of territory. The variation about the line has to be both random and regular. “Regular” here means that the vertical spread is the same as you go from left to right across your scatterplot and that the distribution of points in each vertical slice is roughly normal.

**Yogi:** This is starting to sound complicated.

**Shark:** Not really. Sometimes a violation of the conditions will be obvious from the scatterplot. To be safe—or if you happen to be taking some important test—you should look at a residual plot as well as a dot plot or boxplot of the residuals.

**Yogi:** That still leaves one more condition. Surely you’re not going to tell me I can check randomness by looking at a plot?

**Shark:** No. For that condition, you need to check how the data were collected. The observations should have been selected randomly, which means, partly, that they should have been selected independently.

**Yogi:** *(Loud sigh)*

**Shark:** Just read the next example, and you’ll see how easy it is.
Example: Price Versus Horsepower

In E1 on page 751, you were asked about price versus horsepower for a random sample of car models. Display 11.19 shows the data and a scatterplot. On the face of it, the relationship looks strong enough for you to conclude that the pattern is not simply the result of random sampling: In the population as a whole, there really must be a relationship between price and horsepower. You would expect a formal test to lead to the same conclusion, and it does. Carry out a significance test for the slope of the true regression line for price versus horsepower.

<table>
<thead>
<tr>
<th>Horsepower</th>
<th>Price ($ thousands)</th>
</tr>
</thead>
<tbody>
<tr>
<td>110</td>
<td>13.4</td>
</tr>
<tr>
<td>170</td>
<td>16.3</td>
</tr>
<tr>
<td>165</td>
<td>16.6</td>
</tr>
<tr>
<td>93</td>
<td>11.3</td>
</tr>
<tr>
<td>142</td>
<td>19.0</td>
</tr>
<tr>
<td>214</td>
<td>19.3</td>
</tr>
<tr>
<td>115</td>
<td>14.0</td>
</tr>
<tr>
<td>124</td>
<td>10.0</td>
</tr>
<tr>
<td>225</td>
<td>35.2</td>
</tr>
<tr>
<td>255</td>
<td>32.5</td>
</tr>
<tr>
<td>155</td>
<td>13.5</td>
</tr>
<tr>
<td>200</td>
<td>18.5</td>
</tr>
<tr>
<td>70</td>
<td>8.6</td>
</tr>
<tr>
<td>81</td>
<td>9.1</td>
</tr>
<tr>
<td>168</td>
<td>26.7</td>
</tr>
</tbody>
</table>

Display 11.19  Price versus horsepower for a random sample of 15 car models.

Solution

You have a random sample. The relationship looks reasonably linear in the scatterplot. The equation of the least squares regression line through these points is $\hat{y} = -1.544 + 0.126x$.

If you examine the residual plot in Display 11.20, you can see that while the relationship appears to be generally linear, the variation from the regression line tends to grow with $x$. That is, the values of $y$ tend to fan out and get farther from the regression line as the value of $x$ increases. In the next section, you will see how a transformation helps fix this violation of the conditions for inference.
Check condition: normality.

State the hypotheses.

Do computations and draw a sketch.

Give conclusion in context.

The results in the example are summarized in the Data Desk printout of a regression analysis in Display 11.22. Note that $s$ is the standard deviation of the residuals and is equal to the square root of the mean square for residuals. The
regression coefficients are given in the column labeled “Coefficient.” The estimated standard error of \( b_1 \) is shown in the “Horsepower” line in the column labeled “s.e. of Coeff.” The \( t \)-ratio and its corresponding \( P \)-value finish out this line.

\[
\begin{array}{lllll}
\text{Source} & \text{Sum of Squares} & \text{df} & \text{Mean Square} & \text{F-ratio} \\
\text{Regression} & 663.792 & 1 & 663.792 & 33.5 \\
\text{Residual} & 257.248 & 13 & 19.7883 & \\
\end{array}
\]

Display 11.22  Data Desk summary of the regression of price versus horsepower for a random sample of cars.

Yogi: \( F \)-ratio! Aaargh! See, there is another distribution!

Shark: True, but you can ignore the \( F \)-ratio; here it is equivalent to the \( t \)-test.

As for all statistical tests, there are three ways to get a tiny \( P \)-value like the one in the example: The model could be unsuitable, the sample could be truly unusual, or the null hypothesis could be false.

- **Unsuitable model?** The relationship looks linear, and the sample is random. But we are a bit worried about the fact that the variability of the \( y \)-values seems to grow with the value of \( x \).
- **Unusual sample?** The \( P \)-value tells just how unusual the sample would be if the null hypothesis were true. Here, less than one sample in 10,000 would give such a large value of the test statistic.
- **False \( H_0 \)?** By a process of elimination, this is the most reasonable explanation. What’s the bottom line? In the population as a whole, there’s a positive linear relationship between \( price \) and \( horsepower \).

### DISCUSSION

**Significance Test for a Slope**

D9. Consider the data on phone numbers and names that you collected in Activity 11.2a as having come from a random sample of students. Test to see if there is a linear relationship between the sum of the last four digits of a student’s phone number and the number of letters in his or her full name. Learn to use the \( t \)-test on your calculator to do the computations. [See Calculator Note 11D.] Enter the data into statistical software, if you have it available, and check the computations there as well.

D10. Suppose you have a standard computer output for a regression. What is being tested in the line that includes a \( t \)-statistic for the coefficient of \( x \)? What does the \( P \)-value for this test represent? What is the conclusion if the \( P \)-value is very small, say, 0.001?


**Confidence Interval Estimation**

Now that you know that the price of a car increases with horsepower, your next question might be "How much?" Whenever you reject a null hypothesis that a slope is 0, it is good practice to construct a confidence interval. If the interval is extremely wide, due to large variation in the residuals and small sample size, that tells you that the estimate $b_1$ is practically useless. For example, an estimated increase in annual income of $50 to $10,000 for every additional year of experience doesn't tell you much about what an individual employee's salary increase might be. It also might happen that the slope is so small as to be practically meaningless, even though a large sample size makes it “statistically significant.” For example, a special exercise that lowers your blood pressure by 1 to 1.5 points for every additional 5 hours per week you spend exercising will not attract many takers, even if a huge study proves that the decrease is real. Always temper statistical significance with practical significance.

As in past chapters, a confidence interval takes the form

$$\text{statistic} \pm (\text{critical value}) \cdot (\text{standard deviation of statistic})$$

**Constructing a Confidence Interval for a Slope**

1. **Check conditions.** To get a capture rate equal to the advertised rate, the conditional distributions of $y$ for fixed values of $x$ must be approximately normal, with means that lie on a line and standard deviations that are constant across all values of $x$. Of course, you can't check the population for these; you have to use the sample.
   - Verify that you have a random sample from a bivariate population (or a set of independent random samples, one for each fixed value of $x$), or verify that treatments were assigned randomly to subjects in an experiment.
   - Make a scatterplot and check to see if the relationship looks linear.
   - Make a residual plot to check departures from linearity and that the residuals are of uniform size across all values of $x$.
   - Make a univariate plot of the residuals to see if it's reasonable to assume that they came from a normal distribution.

2. **Do computations.** The confidence interval is

$$b_1 \pm t^* \cdot s_{b_1}$$

The value of $t^*$ depends on the confidence level and the number of degrees of freedom, $df$, which is $n - 2$. [See Calculator Note 11E.]

3. **Give interpretation in context.** For a 95% confidence interval, you would say that you are 95% confident that the slope of the underlying linear relationship lies in the interval. By 95% confidence, you mean that out of every 100 such confidence intervals you construct from random samples, you expect the true value, $\beta_1$, to be in 95 of them.
11.2 Making Inferences About Slopes

Some calculators do not give you the value of $s_b$. However, when testing that $\beta_1 = 0$, they will give you $t$ and $b_1$, so you can compute $s_b$ from

$$ t = \frac{b_1}{s_b} \quad \text{or} \quad s_b = \frac{b_1}{t} $$

**Example: Price Versus Horsepower**

Construct a 95% confidence interval for the slope of the line that relates *price* to *horsepower*.

**Solution**

The conditions are the same as for a test of significance and were checked in the previous example.

You have $15 - 2$, or 13, degrees of freedom, so the value of $t^*$ from Table B on page 826 is 2.160. Thus, a 95% confidence interval estimate of the true slope is given by

$$ b_1 \pm t^* \cdot s_b = 0.126 \pm 2.160(0.0217) = 0.126 \pm 0.047 $$

or $(0.079, 0.173)$.

You can also use a calculator to find this interval. [See Calculator Note 11E.]

You are 95% confident that the slope of the true linear relationship between *price* and *horsepower* is between 0.079 and 0.173. Converting from thousands of dollars to dollars, the increase in the cost of a car per unit increase in horsepower is somewhere between $79 and $173. In other words, if one model has 1 horsepower more than another model, its price tends to be between $79 and $173 more. This result means that any true slope $\beta_1$ between 0.079 and 0.173 could have produced such data as a reasonably likely outcome. A value of $\beta_1$ outside the confidence interval could not have produced numbers like the actual data as a reasonably likely outcome.

**DISCUSSION**

**Confidence Interval Estimation**

D11. Construct 95% confidence intervals for the slopes of these regression lines. State which interval is narrowest and which is widest. Explain why the interval widths differ.

a. for predicting *redness* from *sulfate percentage* for the Mars rocks (use the information in Display 11.11 on page 750)

b. for predicting *redness* from *sulfate percentage* for the Mars soil samples (use the information in Display 11.17 on page 754)

c. for predicting *redness* from *sulfate percentage* for the rocks and soil samples taken together (the fitted slope is 0.6268 and the estimated standard error is 0.11)

D12. Suppose a test of the null hypothesis that the slope of the true regression line is 0 produces a $P$-value of 0.02. What do you know about the 95% confidence interval estimate of the slope?
More About the Degrees of Freedom

Yogi: I was wondering just a teeny little bit about why there are $n - 2$ degrees of freedom.

Shark: Did you read the discussion about $df$ in Section 10.2?

Yogi: That Jodain is pretty sharp! How much sulfur does Jodain have?

Shark: Never mind that. Can you use Jodain’s idea to see why there are $n - 2$ degrees of freedom in the test of significance for a slope?

Yogi: The concept of $df$ should apply here because I am using a sum of squared deviations from the regression line to estimate $\sigma$:

$$\sigma = s = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n - 2}}$$

Shark: Right. The “center” is your regression line. How many deviations, or residuals, are there?

Yogi: There are $n$ values of $(y_i - \hat{y}_i)$, where $n$ is the number of pairs $(x, y)$ in the sample. I suppose two of them must be redundant. If I know $n - 2$ of the residuals, I can figure out the other two?

Shark: Right. Try it with this example, where I won’t tell you either the values of $y$ or the equation of the regression line. The two missing residuals are $R$ and $S$. Can you figure out what they are?

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
<th>$y - \hat{y}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>—?—</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>—?—</td>
<td>-2.5</td>
</tr>
<tr>
<td>3</td>
<td>—?—</td>
<td>$R$</td>
</tr>
<tr>
<td>4</td>
<td>—?—</td>
<td>$S$</td>
</tr>
</tbody>
</table>

Yogi: Well, I know that $\sum(y - \hat{y}) = 0$ because that’s always the case. So $1 + (-2.5) + R + S$ equals 0. But that’s not enough to get $R$ and $S$.

Shark: Correct. You need a second condition on the residuals. That condition is that $x$ and the residuals are uncorrelated—the residuals don’t grow or shrink as $x$ increases. If you check the formula for the correlation, you will see that for it to be zero means that

$$\sum(x - \bar{x})(\text{residual} - \text{mean of residuals}) = \sum(x - \bar{x})(y - \hat{y}) - 0$$

$$= \sum(x - \bar{x})(y - \hat{y}) = 0$$

Yogi: Okay. That gives me two linear equations in two unknowns:

$$\sum(y - \hat{y}) = 1 + (-2.5) + R + S = 0$$

$$\sum(x - \bar{x})(y - \hat{y}) = (1 - 2.5)1 + (2 - 2.5)(-2.5) + (3 - 2.5)R + (4 - 2.5)S = 0$$

The missing residuals are $R = 2$ and $S = -0.5$. 
Shark: Right! So do you see why there are \( n - 2 \) degrees of freedom?

Yogi: Anytime you give me all but two of the residuals, I can find the two missing ones—so they don’t give me any new information about the spread of the points about the regression line. I’ll bet even Jodain doesn’t know that!

**Summary 11.2: Making Inferences About Slopes**

For both a significance test for a slope and a confidence interval for a slope, there are four conditions you should check, which you can remember as brief statements:

- You have a random sample (or a random assignment of treatments to subjects).
- The relationship between the variables looks linear.
- Residuals have equal standard deviations across all values of \( x \).
- Residuals are normal at each fixed \( x \).

You check the first condition by finding out how the observations were collected. Check the second and third conditions by making a scatterplot and residual plot. Check the last condition by making a dot plot or boxplot of the residuals.

As you’ll see in the next section, all is not lost if the last three conditions are not met. But if the observations weren’t selected at random from the population (or, in an experiment, if treatments weren’t randomly assigned), then if you proceed with inference you must state your conclusions very, very cautiously.

To test \( H_0: \beta_1 = 0 \), compute the test statistic

\[
t = \frac{b_1 - 0}{s_{b_1}}
\]

To find the \( P \)-value, compare the value of the test statistic with a \( t \)-distribution with \( n - 2 \) degrees of freedom. If you cannot reject the null hypothesis, then there is no statistically significant evidence of a linear relationship between \( x \) and \( y \).

For a confidence interval for the slope of the true regression line, compute \( b_1 \pm t^* \cdot s_{b_1} \). Again, use \( n - 2 \) degrees of freedom.

As a rule, first do a test to answer the question “Is there an effect?” Then, if you reject \( H_0 \), construct a confidence interval to answer the question “How big is the effect?”

Don’t confuse statistical significance with practical importance. **Significant** means “big enough to be detected with the data available”; **important** means “big enough to care about.”

You should not use the techniques of this section for time-series data, that is, for cases that correspond to consecutive points in time. In these situations, the individual observations typically aren’t selected at random and so are highly dependent. Today’s temperature depends on yesterday’s. The unemployment rate next quarter is unlikely to be very far from the rate this quarter. If your cases have a natural order in time, chances are good that you should use special inference methods for analysis of **time series** rather than the methods of this chapter.
The Test Statistic for a Slope

P10. The regression analysis in Display 11.23 is for the six soil samples from Mars. The value of the test statistic and the corresponding P-value for “Sulfate” are missing. Use only the information in the rest of the printout and Table B to find these values. What is your conclusion?

Display 11.23  Regression analysis for the Mars soil samples.

Dependent variable is: Redness
No Selector
R squared = 3.3% R squared (adjusted) = -20.9%
s = 0.7095 with 6 - 2 = 4 degrees of freedom

Source | Sum of Squares | df | Mean Square | F-ratio
--- | --- | --- | --- | ---
Regression | 0.068487 | 1 | 0.068487 | 0.136
Residual | 2.01345 | 4 | 0.503362 | 

Variable | Coefficient | s.e. of Coeff | t-ratio | prob
--- | --- | --- | --- | ---
Constant | 4.46564 | 2.003 | 2.23 | 0.0897
Sulfate | 0.133399 | 0.3617 | ? | ?

Display 11.24  Sum of phone number digits and number of letters in name.

Significance Test for a Slope

P11. Display 11.24 shows data like those generated in Activity 11.2a on name length and the sum of the last four digits of the phone number for a sample of 25 people, along with the scatterplot and regression line. Comment on the shape of the cloud of points in the scatterplot. Will the t-statistic for testing that the slope is not significantly different from zero be large or small? Explain your reasoning.

P12. For the Mars rocks and soil samples taken together, the fitted slope is 0.6268, with an estimated standard error of 0.11. Use your calculator to compute the t-statistic and find the P-value for testing the hypothesis that the true slope is 0. State your conclusion.

P13. Chirping. Display 11.25 shows the number of times a cricket chirped per second and the air temperature at the time of the chirping. Some people claim that they can tell the temperature by counting cricket chirps.
The regression equation is:

\[ \text{Temp} = 25.2 + 3.29 \times \text{Chirps} \]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>Stdev</th>
<th>t-ratio</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>25.23</td>
<td>10.06</td>
<td>2.51</td>
<td>0.026</td>
</tr>
<tr>
<td>Chirps</td>
<td>3.2911</td>
<td>6.012</td>
<td>5.47</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\[ s = 3.829 \quad R^2 = 69.7\% \quad R^2(\text{adj}) = 67.4\% \]

**Analysis of Variance**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>439.29</td>
<td>439.29</td>
<td>29.97</td>
<td>0.000</td>
</tr>
<tr>
<td>Error</td>
<td>13</td>
<td>190.55</td>
<td>14.66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>629.84</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Display 11.26** Regression analysis for the chirping problem.

P14. Display 11.27 shows IQ and head circumference (in centimeters) for a sample of 20 people. (These data were explored in D21 in Chapter 3.) With head circumference as the predictor and IQ as the response, is there statistically significant evidence of a linear trend?

<table>
<thead>
<tr>
<th>Head Circumference (cm)</th>
<th>IQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>54.7</td>
<td>96</td>
</tr>
<tr>
<td>54.2</td>
<td>89</td>
</tr>
<tr>
<td>53</td>
<td>87</td>
</tr>
<tr>
<td>52.9</td>
<td>87</td>
</tr>
<tr>
<td>57.8</td>
<td>101</td>
</tr>
<tr>
<td>56.9</td>
<td>103</td>
</tr>
<tr>
<td>56.6</td>
<td>103</td>
</tr>
<tr>
<td>55.3</td>
<td>96</td>
</tr>
<tr>
<td>53.1</td>
<td>127</td>
</tr>
<tr>
<td>54.8</td>
<td>126</td>
</tr>
<tr>
<td>57.2</td>
<td>101</td>
</tr>
<tr>
<td>57.2</td>
<td>96</td>
</tr>
<tr>
<td>57.2</td>
<td>93</td>
</tr>
<tr>
<td>57.2</td>
<td>88</td>
</tr>
<tr>
<td>55.8</td>
<td>94</td>
</tr>
<tr>
<td>57.2</td>
<td>85</td>
</tr>
<tr>
<td>57.2</td>
<td>97</td>
</tr>
<tr>
<td>56.5</td>
<td>114</td>
</tr>
<tr>
<td>59.2</td>
<td>113</td>
</tr>
<tr>
<td>58.5</td>
<td>124</td>
</tr>
</tbody>
</table>

Confidence Interval Estimation

P15. Mars rocks contain a high proportion of silicon dioxide (the predominant compound in sand and glass) and much smaller amounts of titanium dioxide (similar to silicon dioxide but with titanium replacing the silicon). Display 11.28 shows a computer printout of the regression of titanium dioxide percentage versus silicon dioxide percentage for the five Mars rocks.

Dependent variable is: \( \% \text{TiO}_2 \)
No Selector
5 total cases
\( R \text{ squared} = 96.1\% \) R squared (adjusted) = 94.8%
\( s = 0.0257 \) with \( 5 - 2 = 3 \) degrees of freedom

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>0.049095</td>
<td>1</td>
<td>0.049095</td>
<td>74.2</td>
</tr>
<tr>
<td>Residual</td>
<td>0.001965</td>
<td>3</td>
<td>0.000662</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>s.e. of Coeff</th>
<th>t-ratio</th>
<th>prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>2.76962</td>
<td>0.2217</td>
<td>12.5</td>
<td>0.0011</td>
</tr>
<tr>
<td>%Si02</td>
<td>–0.033718</td>
<td>0.0039</td>
<td>–8.61</td>
<td>0.0033</td>
</tr>
</tbody>
</table>

Display 11.28 Regression of titanium dioxide percentage versus silicon dioxide percentage for the five Mars rocks.

Exercises

E11. Display 11.30 shows the gas mileage (mpg) and horsepower ratings (hp) for the random sample of car models in E1. The scatterplot and printout for the regression of mpg versus hp are shown in Display 11.31.

<table>
<thead>
<tr>
<th>hp</th>
<th>mpg</th>
</tr>
</thead>
<tbody>
<tr>
<td>110</td>
<td>36</td>
</tr>
<tr>
<td>170</td>
<td>23</td>
</tr>
<tr>
<td>165</td>
<td>20</td>
</tr>
<tr>
<td>93</td>
<td>29</td>
</tr>
<tr>
<td>142</td>
<td>21</td>
</tr>
<tr>
<td>214</td>
<td>28</td>
</tr>
<tr>
<td>115</td>
<td>30</td>
</tr>
<tr>
<td>124</td>
<td>29</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>hp</th>
<th>mpg</th>
</tr>
</thead>
<tbody>
<tr>
<td>225</td>
<td>23</td>
</tr>
<tr>
<td>255</td>
<td>25</td>
</tr>
<tr>
<td>155</td>
<td>31</td>
</tr>
<tr>
<td>200</td>
<td>27</td>
</tr>
<tr>
<td>70</td>
<td>43</td>
</tr>
<tr>
<td>81</td>
<td>33</td>
</tr>
<tr>
<td>168</td>
<td>28</td>
</tr>
</tbody>
</table>

Display 11.30 Gas mileage and horsepower ratings.

a. Locate the estimated standard error of \( b_1 \).
b. Is there evidence to say that the slope of the true regression line is different from 0? Use \( \alpha = 0.05 \).

c. Find a 90% confidence interval estimate of the slope of the true regression line. Interpret the result in the context of the variables.

Dependent variable is: MPG
No Selector
\( R \text{ squared} = 40.6\% \) R squared (adjusted) = 36.0%
\( s = 4.778 \) with \( 15 - 2 = 13 \) degrees of freedom

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>202.759</td>
<td>1</td>
<td>202.759</td>
<td>8.88</td>
</tr>
<tr>
<td>Residual</td>
<td>296.841</td>
<td>13</td>
<td>22.8339</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>s.e. of Coeff</th>
<th>t-ratio</th>
<th>prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>–0.177459</td>
<td>2.289</td>
<td>–0.078</td>
<td>0.9419</td>
</tr>
<tr>
<td>Horsepower</td>
<td>0.026856</td>
<td>0.046046</td>
<td>0.582</td>
<td>0.5916</td>
</tr>
</tbody>
</table>

Display 11.31 Regression analysis for gas mileage versus horsepower rating.
E12. The scatterplot and printout in Display 11.32 show the regression of mean monthly gas usage (in therms) versus mean monthly temperature for a single-family residence over a sample of months. (See Display 11.15 on pages 752–753 for the complete set of data.)

![Scatterplot and computer printout](image)

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>Stdev</th>
<th>t-ratio</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>16.988</td>
<td>1.196</td>
<td>14.20</td>
<td>0.000</td>
</tr>
<tr>
<td>MeanTemp</td>
<td>-0.23643</td>
<td>0.02401</td>
<td>-9.85</td>
<td>0.000</td>
</tr>
</tbody>
</table>

s = 1.548  R·sq = 84.3%  R·sq(adj) = 83.5%

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>232.45</td>
<td>232.45</td>
<td>97.01</td>
<td>0.000</td>
</tr>
<tr>
<td>Error</td>
<td>18</td>
<td>43.13</td>
<td>2.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>19</td>
<td>275.58</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Display 11.32  Scatterplot and computer printout for mean monthly gas usage versus mean monthly temperature.

a. Locate the estimated standard error of the slope.

b. Is there sufficient evidence to say that the slope of the true regression line is different from 0? Use α = 0.05.

c. Find a 90% confidence interval estimate of the slope of the true regression line. Interpret the result in the context of the variables.

E13. Display 11.33 shows the gas mileage (mpg) and maximum revolutions per minute the engine is designed to produce (rpm) for the random sample of car models in E1. Is there evidence that the mean mpg is a linear function of the maximum rpm? Do all four steps of the test of significance.

<table>
<thead>
<tr>
<th>rpm</th>
<th>mpg</th>
</tr>
</thead>
<tbody>
<tr>
<td>5200</td>
<td>36</td>
</tr>
<tr>
<td>4800</td>
<td>23</td>
</tr>
<tr>
<td>4000</td>
<td>20</td>
</tr>
<tr>
<td>4800</td>
<td>29</td>
</tr>
<tr>
<td>5000</td>
<td>21</td>
</tr>
<tr>
<td>5800</td>
<td>28</td>
</tr>
<tr>
<td>5500</td>
<td>30</td>
</tr>
<tr>
<td>6000</td>
<td>29</td>
</tr>
</tbody>
</table>

Display 11.33  Gas mileage and maximum revolutions per minute for the car models problem.

E14. Refer to the data on mean monthly electricity usage (in KWH) and mean monthly temperature (in degrees Fahrenheit) for a single-family residence over a sample of months in Display 11.15 on pages 752–753. Is there evidence that the mean monthly electricity usage is a linear function of the mean monthly temperature?

a. Do all four steps of the test of significance.

b. Would you say that this is strong evidence of a linear relationship or evidence of a strong linear relationship? Why?

E15. Refer to the data on chirp rate in P13.

a. Construct (and interpret, as always) a 95% confidence interval for the slope of the true regression line

i. for predicting the temperature from the chirp rate

ii. for predicting the chirp rate from the temperature

b. Explain why the interval widths in part a are not the same.
E16. Display 11.34 shows the arsenic concentrations in the toenails of a sample of 21 people who used water from their private wells and the arsenic concentration in their well water. Both measurements are in parts per million.

<table>
<thead>
<tr>
<th>Age</th>
<th>Arsenic in Water</th>
<th>Arsenic in Toenails</th>
</tr>
</thead>
<tbody>
<tr>
<td>44</td>
<td>0.00087</td>
<td>0.119</td>
</tr>
<tr>
<td>45</td>
<td>0.00021</td>
<td>0.118</td>
</tr>
<tr>
<td>44</td>
<td>0</td>
<td>0.099</td>
</tr>
<tr>
<td>66</td>
<td>0.00115</td>
<td>0.118</td>
</tr>
<tr>
<td>37</td>
<td>0</td>
<td>0.277</td>
</tr>
<tr>
<td>45</td>
<td>0</td>
<td>0.358</td>
</tr>
<tr>
<td>47</td>
<td>0.00013</td>
<td>0.08</td>
</tr>
<tr>
<td>38</td>
<td>0.00069</td>
<td>0.158</td>
</tr>
<tr>
<td>41</td>
<td>0.00039</td>
<td>0.31</td>
</tr>
<tr>
<td>49</td>
<td>0</td>
<td>0.105</td>
</tr>
<tr>
<td>72</td>
<td>0</td>
<td>0.073</td>
</tr>
<tr>
<td>45</td>
<td>0.046</td>
<td>0.832</td>
</tr>
<tr>
<td>53</td>
<td>0.0194</td>
<td>0.517</td>
</tr>
<tr>
<td>86</td>
<td>0.137</td>
<td>2.252</td>
</tr>
<tr>
<td>8</td>
<td>0.0214</td>
<td>0.851</td>
</tr>
<tr>
<td>32</td>
<td>0.0175</td>
<td>0.269</td>
</tr>
<tr>
<td>44</td>
<td>0.0764</td>
<td>0.433</td>
</tr>
<tr>
<td>63</td>
<td>0</td>
<td>0.141</td>
</tr>
<tr>
<td>42</td>
<td>0.0165</td>
<td>0.275</td>
</tr>
<tr>
<td>62</td>
<td>0.00012</td>
<td>0.135</td>
</tr>
<tr>
<td>36</td>
<td>0.0041</td>
<td>0.175</td>
</tr>
</tbody>
</table>


a. Find (and interpret, as always) a 95% confidence interval estimate of the slope of the true regression line.
   i. for predicting arsenic in toenails from arsenic in well water
   ii. for predicting arsenic in well water from arsenic in toenails
b. Explain why the interval widths in part a are not the same.

c. Choose the data point that you think is exerting the strongest influence on the results. What happens to the answers in part a if this data point is removed?

E17. Refer to the height versus arm span data in Display 11.10 on page 749. Is there evidence in the data to refute Leonardo’s claim that, on average, height is equal to arm span?

E18. When traveling by air, would you rather arrive on time or be relatively sure that your bags would arrive at the same time you do? Or perhaps there is no relationship between these two variables. Use the data in Display 11.35 to test whether there is evidence of a linear relationship that cannot be explained by chance between the percentage of on-time arrivals and the rate of mishandled bags for these major airlines. (These data were explored in Chapter 3.)

<table>
<thead>
<tr>
<th>Airline</th>
<th>Mishandled Bags (per thousand passengers)</th>
<th>Percentage On-Time Arrivals</th>
</tr>
</thead>
<tbody>
<tr>
<td>America West</td>
<td>4.36</td>
<td>81.9</td>
</tr>
<tr>
<td>United</td>
<td>4.00</td>
<td>80.9</td>
</tr>
<tr>
<td>Southwest</td>
<td>4.42</td>
<td>78.4</td>
</tr>
<tr>
<td>US Airways</td>
<td>7.16</td>
<td>78.3</td>
</tr>
<tr>
<td>Continental</td>
<td>4.62</td>
<td>75.7</td>
</tr>
<tr>
<td>JetBlue</td>
<td>5.92</td>
<td>73.8</td>
</tr>
<tr>
<td>American</td>
<td>6.50</td>
<td>73.1</td>
</tr>
<tr>
<td>Delta</td>
<td>8.03</td>
<td>70.1</td>
</tr>
<tr>
<td>Alaska</td>
<td>7.02</td>
<td>69.1</td>
</tr>
<tr>
<td>Northwest</td>
<td>5.36</td>
<td>67.2</td>
</tr>
</tbody>
</table>

Display 11.35 Number of mishandled bags and percentage of on-time arrivals. [Source: U.S. Department of Transportation, Air Travel Consumer Report, October 2005.]

E19. To test the potential effectiveness of laetisaric acid in controlling fungal diseases in crop plants, various concentrations of laetisaric acid were applied to petri dishes containing the fungus Pythium ultimum. After 24 hours, the average radius of the fungus colonies at each concentration was calculated. The data are shown in Display 11.36. Find the
linear regression equation and interpret the slope. Then determine whether the slope is statistically significant.

<table>
<thead>
<tr>
<th>Acid Concentration (μg/ml)</th>
<th>Colony Radius (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>33.3</td>
</tr>
<tr>
<td>0</td>
<td>31.0</td>
</tr>
<tr>
<td>3</td>
<td>29.8</td>
</tr>
<tr>
<td>3</td>
<td>27.8</td>
</tr>
<tr>
<td>6</td>
<td>28.0</td>
</tr>
<tr>
<td>6</td>
<td>29.0</td>
</tr>
<tr>
<td>10</td>
<td>25.5</td>
</tr>
<tr>
<td>10</td>
<td>23.8</td>
</tr>
<tr>
<td>20</td>
<td>18.3</td>
</tr>
<tr>
<td>20</td>
<td>15.5</td>
</tr>
<tr>
<td>30</td>
<td>11.7</td>
</tr>
<tr>
<td>30</td>
<td>10.0</td>
</tr>
</tbody>
</table>


For E20–E22: Display 11.37 shows the selling price of houses, the area of the houses, the number of bedrooms, and the number of bathrooms for a sample of previously owned houses resold in Gainesville, Florida.

E20. Is the slope of the least squares line for predicting price from area statistically significant? If so, interpret the slope.

E21. Fit a least squares regression line to price as a function of number of bedrooms. Interpret the slope of the line. Do you see any weaknesses in this analysis?

E22. Is it appropriate to model price as a linear function of number of bathrooms? Regardless of your answer, fit a regression line and test the significance of the slope. Next, analyze the relationship between these two variables using the techniques of Chapter 9 and compare your two analyses.

<table>
<thead>
<tr>
<th>Price (thousands of dollars)</th>
<th>Area (thousands of square feet)</th>
<th>Number of Bedrooms</th>
<th>Number of Bathrooms</th>
</tr>
</thead>
<tbody>
<tr>
<td>115.0</td>
<td>1.67</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>115.0</td>
<td>2.07</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>69.9</td>
<td>1.52</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>76.0</td>
<td>1.15</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>88.0</td>
<td>1.55</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>100.1</td>
<td>1.85</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>87.9</td>
<td>1.68</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>150.0</td>
<td>2.04</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>107.5</td>
<td>1.85</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>34.8</td>
<td>0.78</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>72.0</td>
<td>1.36</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>98.5</td>
<td>1.51</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>142.5</td>
<td>2.40</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>83.5</td>
<td>1.40</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>68.0</td>
<td>1.45</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>28.0</td>
<td>0.84</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>113.4</td>
<td>1.98</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>65.9</td>
<td>1.22</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>101.9</td>
<td>1.92</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>81.8</td>
<td>1.33</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

Display 11.37  House prices, square footage, and number of bedrooms and bathrooms. [Source: Board of Realtors, Gainesville, Fla., January 1996.]

11.3 Transforming for a Better Fit

Yogi: What are we doing in another section? I already know everything about inference for regression.

Shark: Remember how the last section gave four conditions that are part of the standard regression model?
Yogi: I learned how to check those—verify that it was a random sample plus look at a scatterplot, residual plot, and a dot plot of the residuals. Let’s move on.

Shark: But what do you do if your data don’t meet the conditions?

Yogi: That’s easy. I say the data don’t meet the conditions of inference for regression, and I go on to the next exercise.

Shark: Not so fast. What if you were trying to find a cure for tree blight and those were the only data you had?

Yogi: Rocks don’t get tree blight.

Shark: (Silence)

Yogi: (Mumble/grumble) Okay, what do I do?

Shark: Sometimes, if one or more of the conditions fail to fit your data, you can rescue the situation using the logarithm button on your calculator.

Yogi: Great. I just love that button!

**Checking the Fit of the Model**

More than any other part of inference, model checking is what makes the difference between statistical thinking and mindless number-crunching. Model checking is also the hardest part of inference because often the answers are not clear-cut: A plot shows a hint of curvature at one end; a point is only somewhat apart from its neighbors; a sample is not strictly random, but there are other reasons to think that it is representative. If you are the sort of person who likes definite answers, you might have to push yourself to do justice to model checking.

For inference about the slope of a regression line, there are four conditions to check:

1. You have a random sample or a random assignment of treatments to subjects. (Find out how the data were collected.)
2. The relationship is linear. (Check the scatterplot and residual plot. Sometimes nonlinearity is obvious from the original scatterplot and it is unnecessary to make a residual plot.)
3. Residuals have equal standard deviations across values of \( x \). (Check the residual plot. Again, sometimes it is unnecessary to make a residual plot because it’s obvious from the original scatterplot that the residuals tend to grow or shrink as \( x \) increases.)
4. Residuals are approximately normal at each fixed \( x \). (Make a dot plot or boxplot of only the residuals.)

In some instances, a condition will be clearly violated or clearly satisfied. In others, patterns in the data might merely raise questions without providing clear answers. In still others, there might be so little evidence that there’s not much you
can say. In general, you can be more confident about your judgment when you have a larger number of points than when you have fewer points, but there is no simple rule.

**DISCUSSION**

**Checking the Fit of the Model**

D13. Four scatterplots I–IV, along with their regression lines, residual plots, and dot plots and boxplots of the residuals are shown in Display 11.38.

a. Which plots raise questions about whether the relationship is linear?

b. What features of a plot suggest that the variation in the response is not constant but depends on \( x \)? Which plots show this?

c. What features of a plot suggest that the conditional distributions of \( y \) given \( x \) aren’t normal? Which plots show this?

d. Which plot(s) contains an influential point? Imagine fitting a least squares regression line to this plot using all the data and then removing the influential point and refitting the line. How will the two lines differ?

I. II.

(continued)
Transformations to Improve Linearity

In Display 11.38, you saw both curvature and heteroscedasticity—the points tended to fan out at one end of the scatterplot. You might remember that Chapter 3 gave two basic transformations to straighten curved data. If you have reason to believe the underlying relationship is

- a power function, \( y = ax^b \), try replacing \((x, y)\) with \((\log x, \log y)\)
- an exponential function (either growth or decay), \( y = ab^x \), try replacing \((x, y)\) with \((x, \log y)\)

But what do you do about heteroscedasticity, a common situation with bivariate data? For example, think about the weights of humans plotted against their heights. Almost all babies who are around 20 in. tall will be within a few pounds of each other. However, adults who are 6 ft tall can be hundreds of pounds apart in weight. If you wanted to make an inference about the slope of the regression line in a situation like this, the data would not meet the conditions. As you will see in the example on pages 777–779, a transformation using logarithms often helps with this problem as well as with curvature.
Example: Largemouth Bass

The length of a captured fish is easy for a wildlife biologist to measure with a minimum of handling, but weighing the fish is a bit more difficult. A model to help the biologist predict weight from length would save both time and the health of the fish. Use the data in Display 11.39 to explore a model for the relationship between length and weight of largemouth bass.

Solution

Display 11.39 shows the length and weight measurements for a sample of 11 largemouth bass, along with a scatterplot and a residual plot of these data.

You can see from the plot that the relationship between mean weight and length should not be modeled by the straight line. You can also figure this out by thinking about the physical situation. Length is a linear measure, and weight is more closely connected to volume, a cubic measure. As you learned in Chapter 3, you can linearize power functions by taking the logarithm of each value of $x$ and of each value of $y$. (You can use either base 10 logarithms or natural logarithms for your change of scale.) As you can see from the scatterplot, residual plot, and plot of the residuals in Display 11.40, $\ln(\text{weight})$ versus $\ln(\text{length})$ is linear, and the residuals are small and scattered randomly about the line.
The slope of the regression line for $\ln(\text{length}), \ln(\text{weight})$ is the exponent in the original power function.

Display 11.40 Scatterplot, residual plot, dot plot of residuals, and printout for the least squares regression of $\ln(\text{weight})$ versus $\ln(\text{length})$ for largemouth bass.

Display 11.40 also shows a printout summarizing the regression analysis. The slope, $b_1$, is 3.38051, with an estimated standard error of 0.0353. This is the slope of the regression line for $\ln(\text{weight})$ versus $\ln(\text{length})$:

$$\ln(\text{weight}) = -8.50305 + 3.38051 \cdot \ln(\text{length})$$

Solving for weight

$$\text{weight} = \text{constant} \cdot \text{length}^{3.38051}$$

It appears that the mean weight of largemouth bass is proportional to the length raised to a power that is just a little higher than 3.

Assuming the sample was randomly selected, can you reject the hypothesis that the relationship between length and weight is cubic? With 9 degrees of freedom, the 95% confidence interval for the true slope is given by

$$b_1 \pm t^* \cdot s_{b_1} = 3.38051 \pm (2.262)(0.0353)$$

or (3.301, 3.460). So a slope of 3 is not a plausible value. The relationship is a little bit stronger than cubic.

The slope of the regression line for $(\ln(\text{length}), \ln(\text{weight}))$ is the exponent in the original power function.
**Example: The World’s Women**

This example is based on the data set in Display 11.41, which provides information on population and economic variables for women in a random sample of 30 countries from around the world. The table lists these variables:

- $ps$: percentage of parliamentary seats in single or lower chamber occupied by women
- $fr$: total fertility rate (average number of births per woman)
- $se$: girls’ share of secondary school—middle school and high school—enrollment
- $lew$: life expectancy at birth for women
- $lem$: life expectancy at birth for men
- $img$: infant mortality rate for girls (per 1000 live births)
- $imb$: infant mortality rate for boys (per 1000 live births)

What would be the appropriate model for predicting the variable *infant mortality rate for girls* from *fertility rate*?

**Solution**

The plots in Display 11.42 show the regression of the infant mortality rate for girls ($img$) versus the fertility rate ($fr$). The association is positive and looks as if it follows a linear trend, but the plot is heteroscedastic—countries with larger fertility rates tend to have larger variation in the infant mortality rate for girls. Further, the boxplot of the residuals is skewed left and has four outliers.
Chapter 11 Inference for Regression

Display 11.42 Scatterplot, residual plot, and boxplot of residuals for $\text{img}$ versus $fr$.

Display 11.43 Scatterplots, residual plots, and boxplots of residuals for $\log(\text{img})$ versus $fr$ and $\log(\text{img})$ versus $\log(fr)$.

Compare Display 11.42 with Display 11.43, which shows the same information for the regressions of $\log(\text{img})$ versus $fr$ and $\log(\text{img})$ versus $\log(fr)$. In each case, the transformation has eliminated the heteroscedasticity and the outliers in the residuals by bringing in the larger values and spreading out the smaller ones. The log transformation gives a beautifully symmetrical boxplot of the residuals, but the log-log transformation gives a more randomly scattered residual plot. Either transformation would be acceptable, so the decision of which...
model to use would have to be made on the basis of whether a power model or an exponential model makes more sense in the context. The sample of \( n = 30 \) is a little more than 10% of the total number of countries reporting (about 200), so the formal inference procedures used here would give SEs that are a little too large.

**Yogi:** I have got a great idea! Forget all of this transformation stuff. I just go to the Stat Calc menu on my calculator, fit every function in the list—linear, quadratic, cubic, quartic, log, exponential, power, logistic (whatever that is), and sine—and see which gives me the largest value of \( r \). That’s a whole lot easier than “log this” and “log that.”

**Shark:** Chapter 3 was a long time ago, but . . .

**Yogi:** Oh, right. They wouldn’t let me do that there either. Remind me again why not. After all, \( r \) does tell me how closely the points cluster about my function.

**Shark:** You can get a very high value of \( r \) even though the equation you used isn’t a good fit to the data. Look at Display 11.39 on largemouth bass. It gives a value for \( r^2 \) of 0.95. That tells you that the points cluster closely to the line. But, as you saw, a line isn’t a good model for these data. They clearly follow a curve, not a line. You can see that best from the residual plot—not from the value of \( r \). Also, you can get a very low value of \( r \) even though the points form a fat elliptical cloud and a line is a perfectly appropriate model.

**Yogi:** Well, okay. I see why I shouldn’t pay much attention to \( r \). But if I think the points follow an exponential curve, why can’t I just use my calculator to fit an exponential equation instead of converting all the \( y \)'s to \( \log y \)'s and fitting a straight line? I promise to check the residual plot.

**Shark:** Good statisticians transform for linearity because linear functions are simpler to deal with than curves—and simpler to understand. You know pretty well what the slope and \( y \)-intercept mean for a line, but it's much harder to interpret the parameters for other types of equations.

**Yogi:** Hmmmm. You must be right, because my calculator only tests for the significance of a slope for a line! And the statistical software we are using only fits lines, not other types of functions.

**Shark:** Good point.

---

**DISCUSSION**

**Transformations to Improve Linearity**

D14. Two analyses are shown in Display 11.44. The first shows percentage of renters (\( pr \)) versus total population (\( pop \)) for the 77 largest U.S. cities. The five largest cities were then removed, and the results are shown in the second analysis.
Chapter 11 Inference for Regression

### Analysis with All Cities

Test of Largest U.S. Cities

Response attribute (numeric): PercentRenters  
Predictor attribute (numeric): TotalPopulation  
Sample count: 77

Equation of least-squares regression line:  
\[
\text{PercentRenters} = 2.91679e-06 \times \text{TotalPopulation} + 48.332
\]

Alternative hypothesis: The slope of the least squares regression line is not equal to 0.

The test statistic, Student’s t, is 2.864. There are 75 degrees of freedom (two less than the sample size).

If it were true that the slope of the regression line were equal to 0 (the null hypothesis), and the sampling process were performed repeatedly, the probability of getting a value for Student’s t with an absolute value this great or greater would be 0.0054.

### Analysis with the Five Largest Cities Removed

Test of Largest U.S. Cities-5

Response attribute (numeric): PercentRenters  
Predictor attribute (numeric): TotalPopulation  
Sample count: 72

Equation of least-squares regression line:  
\[
\text{PercentRenters} = 1.64209e-07 \times \text{TotalPopulation} + 49.618
\]

Alternative hypothesis: The slope of the least squares regression line is not equal to 0.

The test statistic, Student’s t, is 0.03653. There are 70 degrees of freedom (two less than the sample size).

If it were true that the slope of the regression line were equal to 0 (the null hypothesis), and the sampling process were performed repeatedly, the probability of getting a value for Student’s t with an absolute value this great or greater would be 0.97.

---

Display 11.44 Analyses of percentage of renters versus total population. [Source: T. Ericson, Data in Depth, CD-ROM (Emeryville, Calif.: Key Curriculum Press, 2001).]

a. Compare the two sets of results. Did eliminating the five largest cities improve the conditions for inference? How influential did the five cities turn out to be?
b. The analysis in Display 11.45 shows a log-log transformation of the data with the five outlying cities removed. Did the log-log transformation further improve the conditions for inference? Did it change the analysis much? What is your conclusion as to the relationship between the percentage of renters and the size of a city?

Log-Log Analysis with the Five Largest Cities Removed

![Graph showing log-log analysis with residual plots and test statistic.]

Test of Largest U.S. Cities-5

<table>
<thead>
<tr>
<th>Response attribute (numeric): logRenters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictor attribute (numeric): logPopulation</td>
</tr>
</tbody>
</table>

Sample count: 72

Equation of least-squares regression line:

\[
\text{logRenters} = 0.0109538 \times \text{logPopulation} + 1.6295
\]

Alternative hypothesis: The slope of the least squares regression line is not equal to 0.

The test statistic, Student’s t, is 0.2548. There are 70 degrees of freedom (two less than the sample size).

If it were true that the slope of the regression line were equal to 0 (the null hypothesis), and the sampling process were performed repeatedly, the probability of getting a value for Student’s t with an absolute value this great or greater would be 0.8.

Display 11.45 Log-log analysis for percentage of renters versus total population.

c. Because these data are for the entire population of the largest U.S. cities, in what sense is the regression line meaningful? In what sense is inference for the slope meaningful?

Summary 11.3: Transforming for a Better Fit

If a scatterplot or residual plot indicates that your data do not fit the conditions for inference very well—the relationship is not linear, the residuals look far from normal, or the variability of y isn’t the same at each fixed x—transforming to a new scale can often put your data in a form that gives a better fit to these conditions.

- If a power model, \( y = ax^b \), fits your data, then the plot of \( \log y \) versus \( \log x \) will be linear. If transforming to logs improves the fit of the linearity assumption, the transformation often will also give residuals that are more nearly normal and variability in \( y \) that is more constant across values of \( x \).
- If an exponential model, \( y = ab^x \), fits your data, then the plot of \( \log y \) versus \( x \) will be linear. There is also a good chance that the transformed data will better fit the other two conditions—that residuals are roughly normal and standard deviations are equal.
Chapter 11 Inference for Regression

Transformations to Improve Linearity

P17. Using the data on largemouth bass from Display 11.39 on page 775 and the regression analysis in Display 11.40, conduct a test of the null hypothesis that the slope of the true regression line is 3 versus the alternative hypothesis that it is not 3.

P18. Suppose a wildlife biologist is working with black crappies instead of largemouth bass, getting the measurements in Display 11.46.

<table>
<thead>
<tr>
<th>Length (in.)</th>
<th>Weight (lb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.4</td>
<td>0.1</td>
</tr>
<tr>
<td>6.6</td>
<td>0.2</td>
</tr>
<tr>
<td>8.4</td>
<td>0.4</td>
</tr>
<tr>
<td>10.0</td>
<td>0.7</td>
</tr>
<tr>
<td>10.5</td>
<td>0.7</td>
</tr>
<tr>
<td>11.1</td>
<td>0.9</td>
</tr>
<tr>
<td>11.4</td>
<td>1.0</td>
</tr>
<tr>
<td>12.0</td>
<td>1.1</td>
</tr>
<tr>
<td>12.5</td>
<td>1.3</td>
</tr>
</tbody>
</table>

Display 11.46 Lengths and weights of black crappies. [Source: Mathematics Teacher 90 (November 1997): 671.]

Practice

Transformations to Improve Linearity

a. Examine the scatterplots, residual plots, and dot plots of the residuals in Display 11.47 for three different models: \((\text{length, weight}), (\text{length, ln(weight)})), and \((\text{ln(length), ln(weight)})\). Give any strengths and weaknesses of each model. Is there any reason to prefer one model over the others?

b. Display 11.48 shows the printout for a significance test that the slope of the regression line for \((\text{ln(length), ln(weight)})\) is significantly different from 0. Use the information to perform a test that the slope is significantly different from 3.

c. Use the printout in Display 11.48 to find a 95% confidence interval for the slope of the transformed data. How does your model differ from the model for largemouth bass?

The regression equation is

\[
\text{LnWeight} = 26.17 + 2.51 \text{LnLength}
\]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>Stdev</th>
<th>t-ratio</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-6.1693</td>
<td>0.2538</td>
<td>-24.31</td>
<td>0.000</td>
</tr>
<tr>
<td>LnLength</td>
<td>2.5115</td>
<td>0.1130</td>
<td>22.22</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Display 11.48 Printout for \((\text{ln(length), ln(weight)})\).
P19. *The World’s Women (continued).* The regression analysis for the logarithm of infant mortality for girls, log(img), versus fertility rate, fr, appears in Display 11.49.

The regression equation is

\[ \text{Log IMG} = 0.679 + 0.195 \text{ FR} \]

29 cases used 1 cases contain missing values

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>Stdev</th>
<th>t-ratio</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.67929</td>
<td>0.09830</td>
<td>6.91</td>
<td>0.000</td>
</tr>
<tr>
<td>FR</td>
<td>0.19502</td>
<td>0.02569</td>
<td>7.59</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\[ s = 0.2588 \text{ R-sq} = 68.1\% \text{ R-sq(adj)} = 66.9\% \]

### Display 11.49  Regression analysis for log(img) versus fr.

a. What would you predict for log(img) for a country that had a fertility rate of five children per woman? Based on these data, describe the distribution of log(img) for countries with fr = 5. What would you predict for img for such a country?

b. Construct a 95% confidence interval for the slope of the regression line. What does this slope indicate about the relationship between the variables fr and img?

P20. Some chimpanzees hunt alone or in small groups, while others hunt in large groups. Not surprisingly, the success of the hunt depends in part on the size of the hunting party. Display 11.50 shows some data on the number of chimps in a hunting party versus the success rate of parties of that size.

### Display 11.50  Hunting party size and percentage successful.

<table>
<thead>
<tr>
<th>Number of Chimps</th>
<th>Percentage Successful</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td>3</td>
<td>28</td>
</tr>
<tr>
<td>4</td>
<td>42</td>
</tr>
<tr>
<td>5</td>
<td>40</td>
</tr>
<tr>
<td>6</td>
<td>58</td>
</tr>
<tr>
<td>7</td>
<td>45</td>
</tr>
<tr>
<td>8</td>
<td>62</td>
</tr>
<tr>
<td>9</td>
<td>65</td>
</tr>
<tr>
<td>10</td>
<td>63</td>
</tr>
<tr>
<td>12</td>
<td>75</td>
</tr>
<tr>
<td>13</td>
<td>75</td>
</tr>
<tr>
<td>14</td>
<td>78</td>
</tr>
<tr>
<td>15</td>
<td>75</td>
</tr>
<tr>
<td>16</td>
<td>82</td>
</tr>
</tbody>
</table>

b. What would you predict for log(img) for a country that had a fertility rate of five children per woman? Based on these data, describe the distribution of log(img) for countries with fr = 5. What would you predict for img for such a country?

b. Construct a 95% confidence interval for the slope of the regression line. What does this slope indicate about the relationship between the variables fr and img?

c. Estimate the slope of the true regression line for the transformed data of part b using a 95% confidence interval. Interpret the result in the context of this problem.

### Exercises

E23. The scatterplot in Display 11.38, part IV, on page 774 shows the crime rate (cr) plotted against the population (pop) for the 76 largest U.S. cities for which data were available. A test of the significance of the slope of the regression line has a P-value of 0.69 with df = 74 and t = -0.4053.

a. Plot the data with size of the hunting party as the explanatory variable. Describe the pattern you see. Will a linear model work well here? Why or why not?

b. Find a transformation that results in a linear regression equation with better predicting ability than a linear equation fit to the original data. What is the slope of your equation?

c. Estimate the slope of the true regression line for the transformed data of part b using a 95% confidence interval. Interpret the result in the context of this problem.

The five largest cities were removed, and Display 11.51 (on the next page) shows analyses with the original scale and with log transformations of both variables.

E24. The scatterplot in Display 11.38, part IV, on page 774 shows the crime rate (cr) plotted against the population (pop) for the 76 largest U.S. cities for which data were available. A test of the significance of the slope of the regression line has a P-value of 0.69 with df = 74 and t = -0.4053.

a. Compare these plots with the plots from the complete set of cities in Display 11.38, part IV. Did eliminating the five
largest cities improve the conditions for inference? Did it change the analysis much? How influential did the five cities turn out to be? Did making a log-log transformation further improve the conditions for inference? Did it change the analysis much? What is your conclusion about the relationship between the variables crime rate and total population?

b. Because these data are for the entire population of largest U.S. cities, in what sense is the regression line meaningful? In what sense is inference for the slope meaningful?

Display 11.51 Scatterplot with regression line, residual plot, and boxplot of residuals for (population, crime rate) and (log(population), log(crime rate)) for U.S. cities with the four largest cities removed.
E24. Are graduation rates for public high schools in the United States related to the ratio of the number of students to the number of teachers? Display 11.52 gives graduation rates and student-teacher ratios for all 50 states. Graduation rate is defined as the percentage of the freshman class from four years earlier who graduated from high school by the year in question.

Suppose you want a model that allows you to predict graduation rate from student-teacher ratio. The plots and tables in Display 11.53 (on the next page) show the analyses for both the original data and the log-transformed data.

a. What conditions for inference are violated by the original data? Did the logarithmic transformation improve those conditions? Did the transformation change the results of the analysis much? What is your conclusion as to the relationship between graduation rates and student-teacher ratios?

b. Do you think any transformation will bring these data into a homogeneous, linear relationship?

c. Because these data comprise the entire population of states, in what sense is the regression line meaningful? In what sense is inference for the slope meaningful?

E25. How does the number of police officers in a state relate to the rate of violent crime?
   a. For the sample of states shown in Display 11.54 (on the next page), find a good-fitting model relating the number of police officers to the violent crime rate.
   b. Construct a 95% confidence interval for the slope and interpret it. (Note that the number of states sampled is greater than 10% of the population, which consists of the 50 states. The only problem caused by that fact is that the estimate of $s_b$ will be a little larger than it should be.)

E26. How do violent crime rates relate to property crime rates in U.S. cities? Display 11.55 (on the next page) gives these rates (in terms of crimes per 100,000 population) for a recent year.
   a. For the sample of cities shown in Display 11.55, find a good-fitting model relating the violent crime rate to the property crime rate.
   b. Construct and interpret a 95% confidence interval estimate of the slope of the true regression line.
Table 11.54 Number of police officers and violent crime rate for a sample of states.

<table>
<thead>
<tr>
<th>State</th>
<th>Officers (thousands)</th>
<th>Violent Crime Rate (per 100,000 population)</th>
</tr>
</thead>
<tbody>
<tr>
<td>California</td>
<td>96.9</td>
<td>622</td>
</tr>
<tr>
<td>Colorado</td>
<td>12</td>
<td>334</td>
</tr>
<tr>
<td>Florida</td>
<td>55.2</td>
<td>812</td>
</tr>
<tr>
<td>Illinois</td>
<td>44.1</td>
<td>657</td>
</tr>
<tr>
<td>Iowa</td>
<td>7.3</td>
<td>266</td>
</tr>
<tr>
<td>Louisiana</td>
<td>16.1</td>
<td>681</td>
</tr>
<tr>
<td>Maine</td>
<td>3.1</td>
<td>110</td>
</tr>
<tr>
<td>Mississippi</td>
<td>8.6</td>
<td>361</td>
</tr>
<tr>
<td>New Jersey</td>
<td>33.4</td>
<td>384</td>
</tr>
<tr>
<td>Tennessee</td>
<td>18.1</td>
<td>707</td>
</tr>
<tr>
<td>Texas</td>
<td>58.9</td>
<td>545</td>
</tr>
<tr>
<td>Virginia</td>
<td>18.8</td>
<td>282</td>
</tr>
<tr>
<td>Washington</td>
<td>14.1</td>
<td>370</td>
</tr>
</tbody>
</table>

Table 11.55 Violent crimes and property crimes in a sample of cities (per 100,000 population).

<table>
<thead>
<tr>
<th>City</th>
<th>Violent Crime Rate</th>
<th>Property Crime Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York, NY</td>
<td>734</td>
<td>2183</td>
</tr>
<tr>
<td>Philadelphia, PA</td>
<td>1378</td>
<td>4175</td>
</tr>
<tr>
<td>San Antonio, TX</td>
<td>598</td>
<td>6844</td>
</tr>
<tr>
<td>Honolulu, HI</td>
<td>288</td>
<td>5336</td>
</tr>
<tr>
<td>Columbus, OH</td>
<td>856</td>
<td>7758</td>
</tr>
<tr>
<td>Memphis, TN</td>
<td>1577</td>
<td>8494</td>
</tr>
<tr>
<td>Boston, MA</td>
<td>1216</td>
<td>4726</td>
</tr>
<tr>
<td>Denver, CO</td>
<td>624</td>
<td>5136</td>
</tr>
<tr>
<td>Washington, DC</td>
<td>1569</td>
<td>5606</td>
</tr>
<tr>
<td>Oklahoma City, OK</td>
<td>890</td>
<td>9123</td>
</tr>
<tr>
<td>Albuquerque, NM</td>
<td>947</td>
<td>6249</td>
</tr>
<tr>
<td>Sacramento, CA</td>
<td>778</td>
<td>6427</td>
</tr>
<tr>
<td>Atlanta, GA</td>
<td>1970</td>
<td>8669</td>
</tr>
<tr>
<td>Miami, FL</td>
<td>1875</td>
<td>6909</td>
</tr>
<tr>
<td>Minneapolis, MN</td>
<td>1193</td>
<td>5390</td>
</tr>
<tr>
<td>Wichita, KS</td>
<td>610</td>
<td>5382</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Flow-Through</th>
<th>Static</th>
</tr>
</thead>
<tbody>
<tr>
<td>23.00</td>
<td>39.00</td>
</tr>
<tr>
<td>22.30</td>
<td>37.50</td>
</tr>
<tr>
<td>9.40</td>
<td>22.20</td>
</tr>
<tr>
<td>9.70</td>
<td>17.50</td>
</tr>
<tr>
<td>0.15</td>
<td>0.64</td>
</tr>
<tr>
<td>0.28</td>
<td>0.45</td>
</tr>
<tr>
<td>0.75</td>
<td>2.62</td>
</tr>
<tr>
<td>0.51</td>
<td>2.36</td>
</tr>
<tr>
<td>28.00</td>
<td>32.00</td>
</tr>
<tr>
<td>0.39</td>
<td>0.77</td>
</tr>
</tbody>
</table>

E27. To measure the effect of certain toxicants found in water, concentrations that kill 50% of the fish in a tank over a fixed period of time (LC50’s) are determined in laboratories. There are two methods for conducting these experiments. One method has water continuously flowing through the tanks, and the other has static water conditions. The Environmental Protection Agency (EPA) wants to adjust all results to the flow-through conditions. Given the data in Display 11.56 on a sample of ten toxicants, establish a model that will allow adjustment of the static values to corresponding flow-through values.

E28. Metals, especially copper and its alloys, are often made by heating (sintering) powders of essentially pure ore. As the generally round particles of powder soften and merge, the spaces (voids) between them become smaller. The volume of the voids per unit volume of the metal is a measure of the porosity of the sintered metal, and porosity is related to the strength of the metal. Porosity is measured by the weight of liquid wax that is taken up by the sintered metal.

The data in Display 11.57 (on the next page) are from a process for sintering copper and tin powders to make bronze bearings. Model the relationship between the sintering time and the weight of wax taken up by the bearings. Construct a confidence interval for the slope of the true regression line and interpret it.
### Display 11.57
Sintering time and weight of wax from a sintering process. [Source: R. Sheaffer and James McClave, Probability and Statistics for Engineers (Boston: Duxbury Press, 1995), p. 536.]

<table>
<thead>
<tr>
<th>Sintering Time (min)</th>
<th>Wax Weight (g)</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>0.615</td>
</tr>
<tr>
<td>7</td>
<td>0.606</td>
</tr>
<tr>
<td>7</td>
<td>0.611</td>
</tr>
<tr>
<td>9</td>
<td>0.586</td>
</tr>
<tr>
<td>11</td>
<td>0.511</td>
</tr>
<tr>
<td>11</td>
<td>0.454</td>
</tr>
<tr>
<td>11</td>
<td>0.440</td>
</tr>
<tr>
<td>13</td>
<td>0.393</td>
</tr>
<tr>
<td>15</td>
<td>0.322</td>
</tr>
<tr>
<td>15</td>
<td>0.343</td>
</tr>
<tr>
<td>15</td>
<td>0.341</td>
</tr>
</tbody>
</table>

### E29.
At the age of 31, Galileo conducted a series of experiments on the path of projectiles that eventually led to his formulation of theories for the motion of falling bodies. In Experiment 1, a ball was released at a set height on an inclined ramp and allowed to roll down a groove set into the ramp. After leaving the ramp, the ball fell to the floor unobstructed. The measurement of the release height ($H$) and the horizontal distance traveled ($D$) between leaving the ramp and hitting the floor were recorded, measured in punti.

Realizing that the ball leaving the ramp in the first experiment had a downward velocity when it left the ramp, Galileo carried out Experiment 2, in which a narrow horizontal shelf was placed at the end of the ramp. When the ball reached the edge of the inclined plane, it rolled across the shelf before starting its fall, neutralizing the downward force. The data for both experiments are shown in Display 11.58.

<table>
<thead>
<tr>
<th>Experiment 1</th>
<th>Experiment 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance 1</td>
<td>Distance 2</td>
</tr>
<tr>
<td>Height 1</td>
<td>Height 2</td>
</tr>
<tr>
<td>573</td>
<td>1500</td>
</tr>
<tr>
<td>534</td>
<td>1340</td>
</tr>
<tr>
<td>495</td>
<td>1328</td>
</tr>
<tr>
<td>451</td>
<td>1172</td>
</tr>
<tr>
<td>395</td>
<td>800</td>
</tr>
<tr>
<td>337</td>
<td>300</td>
</tr>
<tr>
<td>253</td>
<td>800</td>
</tr>
</tbody>
</table>

### Display 11.58

### E30.
Rivers and streams carry sediment (small particles of rock and mineral) downhill as they flow. It seems reasonable that fast-moving streams would carry larger particles than would slower-moving streams. Knowing the relationship between the speed of the water and the size of the sediment particles would be of great value to those studying, for example, the effects of dikes and buildings on stream flow and the resulting sediment carried off by the stream. Display 11.59 shows the sample data on the diameters of particles moved and the speed of the water moving them.

<table>
<thead>
<tr>
<th>Distance (m/s)</th>
<th>Particle Diameter (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.5</td>
</tr>
<tr>
<td>4</td>
<td>0.6</td>
</tr>
<tr>
<td>5</td>
<td>0.7</td>
</tr>
<tr>
<td>6</td>
<td>0.8</td>
</tr>
<tr>
<td>7</td>
<td>0.9</td>
</tr>
<tr>
<td>8</td>
<td>1.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Distance (m/s)</th>
<th>Particle Diameter (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.4</td>
</tr>
<tr>
<td>4</td>
<td>0.5</td>
</tr>
<tr>
<td>5</td>
<td>0.6</td>
</tr>
<tr>
<td>6</td>
<td>0.7</td>
</tr>
<tr>
<td>7</td>
<td>0.8</td>
</tr>
<tr>
<td>8</td>
<td>0.9</td>
</tr>
</tbody>
</table>

### E29.

a. For Experiment 2, set up a plot to predict distance from height. Would a simple linear model be a good fit for these data? If not, find a transformation that will linearize the relationship.

b. Fit a linear equation to the transformed data of part a and estimate the slope in a 95% confidence interval. Interpret the observed slope and the confidence interval estimate.

c. Now plot the data from Experiment 1 and try the same transformation that you used in part b. Does this transformation linearize the relationship here? Does the shelf appear to make a difference in the relationship between height and distance?

### E30.

a. Set up a plot to predict the size of objects moved from the speed of the current. Is a linear model a good fit? If not, find a transformation that linearizes the relationship.

b. Fit a line to the transformed data of part a and estimate the slope in a 95% confidence interval. Interpret the observed slope and the confidence interval.
<table>
<thead>
<tr>
<th>Diameter of Objects Moved (mm)</th>
<th>Speed of Current (m/s)</th>
<th>Classification of Objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.10</td>
<td>Mud</td>
</tr>
<tr>
<td>1.3</td>
<td>0.25</td>
<td>Sand</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>Gravel</td>
</tr>
<tr>
<td>11</td>
<td>0.75</td>
<td>Coarse gravel</td>
</tr>
<tr>
<td>20</td>
<td>1.00</td>
<td>Pebbles</td>
</tr>
<tr>
<td>45</td>
<td>1.50</td>
<td>Small stones</td>
</tr>
<tr>
<td>80</td>
<td>2.50</td>
<td>Large stones</td>
</tr>
<tr>
<td>180</td>
<td>3.50</td>
<td>Boulders</td>
</tr>
</tbody>
</table>

**Display 11.59**  Diameters of particles and the speed of moving water. [Source: www.seattlecentral.edu.]

E31. Data on the world’s women can be found in Display 11.41 on page 777. Display 11.60 shows a regression analysis of the relationship between fertility rate and life expectancy.

\[ \text{Fertility Rate} = -0.140 \cdot \text{Life Expectancy} + 13; r^2 = 0.68 \]

a. Do the data appear to meet the conditions for a regression analysis?
b. Verify the computations in the test of significance for the slope in Display 11.60. Interpret the result of the test.
c. Can you infer that an increase in life expectancy causes the fertility rate to decrease?
d. Does the scatterplot suggest a different way of looking at these data?

E32. Display 11.61 shows a regression analysis of the percentage of parliamentary seats in a single or lower chamber occupied by women (ps) versus the girls’ share of secondary school enrollment (se) from the world’s women data on page 777. Perform a significance test for the slope. Does it agree with the results shown? Make the necessary dot plot of the residuals by estimating them from the plots. Interpret your results.

\[ ps = 0.633 \cdot se - 16; r^2 = 0.14 \]

**Test of World’s Women**

**Response attribute (numeric): PS**
**Predictor attribute (numeric): SE**

<table>
<thead>
<tr>
<th>Sample count: 19</th>
<th>Equation: $PS = 0.633344 \cdot SE - 15.899$</th>
</tr>
</thead>
</table>

Ho: Slope = 0
Ha: Slope is not equal to 0

Student’s $t$: 1.65
DF: 17
P-value: 0.12

**Display 11.61**  Regression analysis of $ps$ versus $se$.
Chapter Summary

When two variables are both quantitative, you can display their relationship in a scatterplot, and you might be able to summarize that relationship by fitting a line to the original data or to the data after a transformation. Although many relationships are more complicated than this, linear relationships occur often enough that many questions can be recast as questions about the slope of a fitted line.

That is old news. What is new in this chapter is that you can use the \( t \)-statistic to test whether the slope of the true regression line is 0. If you can reject that possibility, your regression line will be useful in predicting \( y \) given \( x \). If you can’t, then it’s plausible that the positive or negative trend you see is due solely to the chance variation that always results when you have only a sample.

You also have learned to find a confidence interval for the slope of the true regression line, which helps you decide whether the linear relationship is statistically significant and whether it has any practical significance.

One way to put the methods of this chapter into a larger picture is to remind yourself of the inference methods you’ve learned so far: for one proportion and for the difference of two proportions (Chapter 8), for one mean and for the difference between two means (Chapter 9), for the relationship between two categorical variables (Chapter 10), and now for the relationship between two quantitative variables. In the next chapter, you will work on four case studies that bring all these ideas together.

**Review Exercises**

E33. Study the scatterplots in Display 11.62.

I.

```
   y
3  4  5  6  7  8
---
x  3  4  5  6  7  8
```

II.

```
   y
30  60  90
---
x  0  10  20  30  40  50  60  70
```

III.

```
   y
2  3  4  5  6  8
---
x  3.0  4.0  5.0  6.0  7.0  8.0
```

IV.

```
   y
2  4  6
---
x  2  3  4  5  6  7  8
```

V.

```
   y
2  4  6  8
---
x  2  3  4  5  6  7  8  9  10
```

**Display 11.62** Five scatterplots.

- In which scatterplots is it reasonable to model the relationship between \( y \) and \( x \) with a straight line?
- If you fit a line through each scatterplot by the method of least squares, which plot will give a line with slope closest to 0?
- Which plot shows a correlation coefficient closest to 1?
- For each scatterplot that does not look as if it should be modeled by a straight line, suggest a way to modify the data to make the shape of the plot more nearly linear.

E34. More on pesticides in the Wolf River.

Display 11.63 shows four scatterplots of \( HCB \) concentration versus \( aldrin \) concentration. The four scatterplots are for the measurements taken on the bottom, at mid-depth, at the surface, and for all three locations together. Based on the plots, which depth do you expect to give the narrowest
confi dence interval for the slope? The widest? Give your reasoning.


E35. “If I change to a brand of pizza with lower fat, will I also reduce the number of calories?” Display 11.64 shows the printout for a significance test of calories versus fat per 5-oz serving for 17 popular brands of pizza.

Calories = 241 + 7.26 Fat
Predictor Coef Stdev t-ratio p
Constant 240.55 14.88 16.17 0.000
Fat 7.263 1.064 6.82 0.000
s = 14.36 R-sq = 75.6% R-sq(adj) = 74.0%
Analysis of Variance
Source DF SS MS F p
Regression 1 9605.9 9605.9 46.55 0.000
Error 15 3095.2 206.3
Total 16 12701.1

Display 11.64  Regression analysis for pizza data.

a. Use Display 11.64 to estimate, in a 95% confidence interval, the reduction in calories you can expect per 1-g decrease in the fat content of the pizza.
b. Use your answer to part a to estimate, in a 95% confidence interval, the reduction in calories you can expect per 5-g decrease in the fat content of the pizza.

e36. As in all wars, espionage played a critical role in the American Civil War. One of the more interesting examples concerns the role of detective Allan Pinkerton in using spying methods to estimate the number of troop regiments in the various states. Display 11.65 shows his estimates for September 1861, along with the actual number as discovered after the war.

<table>
<thead>
<tr>
<th>State</th>
<th>Actual Number of Regiments</th>
<th>Pinkerton’s Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alabama</td>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td>Arkansas</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Florida</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Georgia</td>
<td>22</td>
<td>18</td>
</tr>
<tr>
<td>Kentucky</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Louisiana</td>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td>Maryland</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Mississippi</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>North Carolina</td>
<td>15</td>
<td>14</td>
</tr>
<tr>
<td>South Carolina</td>
<td>9</td>
<td>18</td>
</tr>
<tr>
<td>Tennessee</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Texas</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Virginia</td>
<td>50</td>
<td>45</td>
</tr>
</tbody>
</table>


a. Plot the data, with Pinkerton’s number as the explanatory variable. Does the plot have a definite linear trend? If so, fit a least squares regression line to the data.
b. Although these observational data cover all states with Confederate regiments, inference might still be meaningful in deciding if a slope of this size could have happened merely by chance. Conduct a test to see if the observed slope was likely to occur by chance. What is your conclusion?
c. Interpret the slope of the least squares line in part a. Estimate plausible values for the “true” slope in a 95% confidence interval. If Pinkerton was on target with...
his estimates, what value would you get for
the slope of the line? Is this value in
the confidence interval?

d. Virginia contributed by far the most
troops to the Confederate war effort.
What influence does Virginia have on the
analysis in parts a–c?

E37. How are film ratings associated with film
lengths, and are either of these variables
associated with the year of the film’s release?
The data in Display 11.66 are a random
sample of 20 films (TV films excluded) from
the 1996 Movie and Video Guide by Leonard
Maltin, which includes over 19,000 films.
Each movie is rated on the scale 1, 1.5, 2, 2.5,
3, 3.5, 4, with 4 being excellent.

<table>
<thead>
<tr>
<th>Title</th>
<th>Year</th>
<th>Length (min)</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car Wash</td>
<td>1976</td>
<td>97</td>
<td>2.5</td>
</tr>
<tr>
<td>Jason’s Lyric</td>
<td>1994</td>
<td>119</td>
<td>2</td>
</tr>
<tr>
<td>Hold Back Tomorrow</td>
<td>1955</td>
<td>75</td>
<td>1.5</td>
</tr>
<tr>
<td>Mad Love</td>
<td>1995</td>
<td>95</td>
<td>1.5</td>
</tr>
<tr>
<td>Conflict</td>
<td>1945</td>
<td>86</td>
<td>2.5</td>
</tr>
<tr>
<td>It Came from Outer Space</td>
<td>1953</td>
<td>81</td>
<td>3</td>
</tr>
<tr>
<td>And the Ship Sails On</td>
<td>1984</td>
<td>138</td>
<td>3</td>
</tr>
<tr>
<td>Five Golden Hours</td>
<td>1961</td>
<td>90</td>
<td>2</td>
</tr>
<tr>
<td>Brewster McCloud</td>
<td>1970</td>
<td>101</td>
<td>3</td>
</tr>
<tr>
<td>Calling Philo Vance</td>
<td>1940</td>
<td>62</td>
<td>2</td>
</tr>
<tr>
<td>Lady Dracula</td>
<td>1973</td>
<td>80</td>
<td>1</td>
</tr>
<tr>
<td>Evergreen</td>
<td>1934</td>
<td>90</td>
<td>3</td>
</tr>
<tr>
<td>Action in the North Atlantic</td>
<td>1943</td>
<td>127</td>
<td>3</td>
</tr>
<tr>
<td>A Ticklish Affair</td>
<td>1963</td>
<td>89</td>
<td>2</td>
</tr>
<tr>
<td>Four Jills in a Jeep</td>
<td>1944</td>
<td>89</td>
<td>2.5</td>
</tr>
<tr>
<td>Blaze</td>
<td>1989</td>
<td>119</td>
<td>2.5</td>
</tr>
<tr>
<td>Hitler—Dead or Alive</td>
<td>1943</td>
<td>70</td>
<td>2</td>
</tr>
<tr>
<td>Benson Murder Case</td>
<td>1930</td>
<td>69</td>
<td>2.5</td>
</tr>
<tr>
<td>City Lights</td>
<td>1985</td>
<td>85</td>
<td>1</td>
</tr>
<tr>
<td>Galileo</td>
<td>1973</td>
<td>145</td>
<td>3</td>
</tr>
</tbody>
</table>

Display 11.66 Film lengths and ratings. [Source:
Thomas L. Moore, “Paradoxes in Film Ratings,”

a. Is there a statistically significant linear
relationship between year and length?
Remember to do all parts of the analysis,
including the construction of appropriate
plots.

b. Is there a statistically significant
relationship between length and rating?
Again, do all parts of the analysis.

c. Given the results from parts a and b, what
do you think the relationship between
year and rating will be? Is this, in fact,
the case? Make a plot and conduct a
significance test of a slope as part of your
answer.

d. Display 11.67 shows a plot of ratings
by year, separated by short versus long
movies, with “short” being defined as less
than 90 minutes running length. (In the
plot, a circle represents “short.”) Do the
decreasing trends over the years within
the length groups appear to be stronger
than the decreasing trend in the overall
sample not accounting for length? Test
for significance of these two slopes and
compare the results to your answer in
part c.

[Graph showing ratings by year, separated by short
and long movies.]

Display 11.67 Ratings by year, separated by short
and long movies.
E38. The data in Display 11.68 extend the data on the number of police officers and the violent crime rate in Display 11.54 on page 787 and come from a random sample of the 50 U.S. states. The variable \textit{expenditures} is the amount (in millions of dollars) it takes to keep the given number of police officers on the job for 1 year, \textit{population} is the population of the state in millions of people.

\begin{tabular}{lccc}
\hline
State & Officers & Expenditures & Population \\
& (thousands) & ($ millions) & (millions) \\
\hline
California & 96.9 & 7653 & 34 \\
Colorado & 12 & 753 & 4.3 \\
Florida & 55.2 & 3371 & 16 \\
Illinois & 44.1 & 2718 & 12.4 \\
Iowa & 7.3 & 346 & 2.9 \\
Louisiana & 16.1 & 635 & 4.5 \\
Maine & 3.1 & 118 & 1.3 \\
Mississippi & 8.6 & 337 & 2.8 \\
New Jersey & 33.4 & 1829 & 8.4 \\
Tennessee & 18.1 & 828 & 5.7 \\
Texas & 58.9 & 2866 & 20.9 \\
Virginia & 18.8 & 965 & 7.1 \\
Washington & 14.1 & 854 & 5.9 \\
\hline
\end{tabular}


a. How many new police officers would you expect could be added for an increase of $1 million in expenditure?

b. Does removal of the most influential point in the set have much of an effect on your answer to part a?

c. If one state has 1 million more people than another state, how many more police officers would the first state be expected to have?

d. Does removing the most influential point have much of an effect on your answer to part c?

E39. In a bivariate population, what is the relationship between the standard deviation of the responses and the standard deviation of the prediction errors?

E40. The data in Display 11.69 represent nearly 200 years of earthquake activity in the New Madrid area of southeastern Missouri and adjacent states. The magnitude of the earthquakes is based on the Richter scale. The frequency of each magnitude is expressed as the cumulative relative frequency of earthquakes of magnitude greater than or equal to that size. (For example, about 77.6% of the recorded earthquakes were of magnitude 1.66 or greater.) A good model for relating magnitude to cumulative relative frequency would be of help to scientists who study earthquakes. Can you find such a model and justify it statistically?

\begin{tabular}{lcc}
\hline
Magnitude & Cumulative Relative Frequency \\
\hline
1.66 & 77.62 \\
1.75 & 67.61 \\
1.84 & 52.48 \\
1.95 & 40.74 \\
2.10 & 33.11 \\
2.12 & 26.91 \\
2.22 & 20.89 \\
2.33 & 14.12 \\
2.50 & 12.02 \\
2.57 & 10.96 \\
2.67 & 8.71 \\
2.78 & 6.16 \\
2.85 & 4.90 \\
2.98 & 4.07 \\
3.05 & 3.63 \\
3.19 & 3.09 \\
3.32 & 1.44 \\
3.40 & 2.04 \\
3.52 & 1.58 \\
\hline
\end{tabular}

Display 11.69 \textit{Earthquake magnitude and cumulative relative frequency on the New Madrid fault.} \[ \text{Source: www.seattlecentral.edu.} \]
AP1. A statistics exam has two parts, free response and multiple choice. A regression equation for predicting the score, \( f \), on the free response part, from the score, \( m \), on the multiple choice part is \( f = 50 + 0.25m \). This equation was based on the scores of 17 students. The standard deviation of their multiple choice scores was 30, the standard deviation of their free response scores was 16, and the sum of the squared residuals was 2940. What is the estimated standard error of the slope?

- 0.117
- 0.133
- 0.219
- 0.467
- 3.5

AP2. Which of the following is not an important condition to check before constructing a confidence interval for the slope of the true regression line?

- You have a simple random sample.
- The points fall in an elliptical cloud.
- The residuals for small values of \( x \) have about the same variability as the residuals for large values of \( x \).
- The sum of the squared residuals is small.
- The residuals are approximately normally distributed.

AP3. On the first day of statistics class, data on shoe size and height were gathered for 82 randomly selected female high school seniors. Part of a regression analysis is given below. Which of the following is the best interpretation of the \( P \)-value for shoe size?

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coefficient</th>
<th>Std Error</th>
<th>t statistic</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>55.4174</td>
<td>1.1681</td>
<td>47.443</td>
<td>0.0000</td>
</tr>
<tr>
<td>ShoeSize</td>
<td>1.2116</td>
<td>0.1438</td>
<td>8.425</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

\[ s = 1.87862 \]

- An error has been made in the data collection or analysis.
- There is statistically significant evidence of a non-zero slope in the true linear relationship between shoe size and height.
- There isn’t statistically significant evidence of a non-zero slope in the true linear relationship between shoe size and height.

AP4. Refer to AP3. What is the best explanation of what is measured by \( s = 1.87862 \)?

- the variability in the slope from sample to sample
- the variability in the \( y \)-intercept from sample to sample
- the variability in the shoe sizes
- the variability in the heights
- the variability in the residuals

AP5. Refer to AP3. What is the best explanation of what is measured by \( \text{Std Error} = 0.1438 \)?

- the variability in the slope from sample to sample
- the variability in the \( y \)-intercept from sample to sample
- the variability in the shoe sizes
- the variability in the heights
- the variability in the residuals

AP6. Refer to AP3. Which of the following is the appropriate computation for a 95% confidence interval for the slope?

\[ 1.2116 \pm 1.990 \cdot 1.87862 \]
\[ 1.2116 \pm 1.990 \cdot 0.1438 \]
\[ 1.2116 \pm 1.96 \cdot 0.1438/\sqrt{80} \]
\[ 1.2116 \pm 1.990 \cdot 0.1438/\sqrt{80} \]
\[ 1.2116 \pm 1.990 \cdot 1.87862/\sqrt{80} \]

AP7. In an attempt to predict adult heights, researchers randomly selected men and collected their heights at age two and their adult heights, then computed a least squares regression equation, \( \text{adult height} = 1.9 \cdot \text{height at age two} + 3.5 \). A 95% confidence interval for the slope was given as (1.8, 2.0). Which of the following is the best interpretation of this confidence interval?
If the researchers took 100 more random samples, they would expect 95 of the regression equations to have a slope between 1.8 and 2.0.

95% of two-year-old boys have heights between 1.8 and 2.0.

The researchers are pretty sure that if they studied all men, the slope of the true regression line would be between 1.8 and 2.0.

If a two-year-old boy’s height is known, there’s a 95% chance that his adult height will be between 1.8 and 2.0 times his current height.

In predicting the height of an adult man using his height at age two, 95% of the errors will be between 1.8 and 2.0 inches.

AP8. A friend is doing a regression analysis to predict the mass of a small bag of fries given the number of fries in the bag. He finds a linear regression equation and then makes the following residual plot. He asks your advice about whether to proceed with a test of significance of the slope. Which is the best advice you could give?

- A: Something’s wrong here. This can’t be a residual plot.
- B: The relationship is linear. It’s okay to proceed with the test of significance.
- C: The relationship doesn’t have constant variability across all values of $x$. It’s not okay to proceed.
- D: Try a transformation before proceeding.
- E: It’s clear even without a significance test that the true slope isn’t 0.

Investigative Tasks

AP9. Do animals make optimal decisions? Tim Penning of Hope College studied the strategy his dog, Elvis, uses in retrieving a ball thrown into the water of Lake Michigan. Looking at the diagram in Display 11.70, suppose Tim and Elvis stand on the edge of the lake at $A$ and the ball is thrown to $B$ in the lake. Elvis could jump into the water immediately at $A$ and swim to the ball, but he seems to know that he can run faster than he can swim. He could run all the way to $C$ and then swim the perpendicular distance to the ball, but that is not the most time-efficient strategy either. What Elvis actually does is run to a point $D$ and then swim diagonally to the ball at $B$. But, does he determine $D$ so as to minimize the time it takes him to get to the ball?

Methods of calculus can be used to show that the time is minimized when $y$ is related to $x$ by the formula

$$y = \frac{x}{\sqrt{r/s} + 1} \sqrt{r/s} - 1$$
where \( r \) is Elvis's running speed and \( s \) is his swimming speed. Further experiments with Elvis determined that his running speed was about 6.40 m/s and his swimming speed was about 0.91 m/s, which result in the optimal relationship for minimizing time to the ball of

\[
y = 0.144x
\]

Display 11.71 shows the measurements \( x \) and \( y \) (in meters) for 35 ball-tossing experiences with Elvis.

<table>
<thead>
<tr>
<th>( x )</th>
<th>( y )</th>
<th>( x )</th>
<th>( y )</th>
<th>( x )</th>
<th>( y )</th>
<th>( x )</th>
<th>( y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.5</td>
<td>2.0</td>
<td>6.6</td>
<td>1.0</td>
<td>13.5</td>
<td>1.8</td>
<td>12.5</td>
<td>1.5</td>
</tr>
<tr>
<td>7.2</td>
<td>1.0</td>
<td>14.0</td>
<td>2.6</td>
<td>14.2</td>
<td>1.9</td>
<td>15.3</td>
<td>2.3</td>
</tr>
<tr>
<td>10.3</td>
<td>1.8</td>
<td>13.4</td>
<td>1.5</td>
<td>14.2</td>
<td>2.5</td>
<td>11.8</td>
<td>2.2</td>
</tr>
<tr>
<td>11.7</td>
<td>1.5</td>
<td>6.5</td>
<td>1.0</td>
<td>10.9</td>
<td>2.2</td>
<td>7.5</td>
<td>1.4</td>
</tr>
<tr>
<td>12.2</td>
<td>2.3</td>
<td>11.8</td>
<td>2.4</td>
<td>11.2</td>
<td>1.3</td>
<td>11.5</td>
<td>2.1</td>
</tr>
<tr>
<td>19.2</td>
<td>4.2</td>
<td>4.7</td>
<td>0.9</td>
<td>15.0</td>
<td>3.8</td>
<td>12.7</td>
<td>2.3</td>
</tr>
<tr>
<td>11.4</td>
<td>1.3</td>
<td>11.6</td>
<td>2.2</td>
<td>14.5</td>
<td>1.9</td>
<td>6.6</td>
<td>0.8</td>
</tr>
<tr>
<td>17.0</td>
<td>2.1</td>
<td>11.5</td>
<td>1.8</td>
<td>6.0</td>
<td>0.9</td>
<td>15.3</td>
<td>3.3</td>
</tr>
<tr>
<td>15.6</td>
<td>3.9</td>
<td>9.2</td>
<td>1.7</td>
<td>14.5</td>
<td>2.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Describe the pattern in the scatterplot.
b. Assuming this to be a random sample of Elvis's work, construct a 90% confidence interval estimate of the slope of the true regression line relating distances \( x \) and \( y \). Does it look like Elvis is making an optimal decision?
c. Assuming this to be a random sample of Elvis's work, construct a test of the hypothesis, at the 0.05 significance level, that the slope of the true regression line relating distances \( x \) and \( y \) is the optimal 0.144. Now does it look like Elvis is making an optimal decision?
d. Choose the point that you think is most influential on the slope of the regression line and the correlation. Describe what will happen to the slope of the regression line and the correlation if it is removed and the data re-analyzed.
e. The analysis below shows that a regression model fit to the data in Display 11.71 does not have a \( y \)-intercept of 0, as the optimal model does. Explain what will happen to the slope of the regression line shown in the plot if you force the line to go through the origin.

The regression equation is

\[
y = -0.328 + 0.196x
\]

\[
\text{Preditor} \quad \text{Coeff} \quad \text{Stdev} \quad \text{t-ratio} \quad \text{p}
\]

| Constant | -0.3277 | 0.3207 | -1.02 | 0.314 |
| \( x \)  | 0.19647 | 0.02625| 7.49  | 0.000 |

s = 0.5162 \quad R^2 = 62.9\% \quad R^2(adj) = 61.8\%

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>14.935</td>
<td>14.935</td>
<td>56.04</td>
<td>0.000</td>
</tr>
<tr>
<td>Error</td>
<td>33</td>
<td>8.795</td>
<td>0.267</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>23.730</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Display 11.72 Analysis of Elvis’s data.
AP10. Display 11.73 shows birthrate versus gross national product for a sample of countries.

<table>
<thead>
<tr>
<th>Country</th>
<th>Live Births (per 1000 population)</th>
<th>GNP (thousands of dollars per capita)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algeria</td>
<td>29.0</td>
<td>1.6</td>
</tr>
<tr>
<td>Argentina</td>
<td>19.5</td>
<td>4.0</td>
</tr>
<tr>
<td>Australia</td>
<td>14.1</td>
<td>16.6</td>
</tr>
<tr>
<td>Brazil</td>
<td>21.2</td>
<td>2.6</td>
</tr>
<tr>
<td>Canada</td>
<td>13.7</td>
<td>20.8</td>
</tr>
<tr>
<td>China</td>
<td>17.8</td>
<td>1.3</td>
</tr>
<tr>
<td>Cuba</td>
<td>14.5</td>
<td>1.6</td>
</tr>
<tr>
<td>Denmark</td>
<td>12.4</td>
<td>24.2</td>
</tr>
<tr>
<td>Egypt</td>
<td>28.7</td>
<td>0.5</td>
</tr>
<tr>
<td>France</td>
<td>13.0</td>
<td>24.1</td>
</tr>
<tr>
<td>Germany</td>
<td>11.0</td>
<td>19.8</td>
</tr>
<tr>
<td>India</td>
<td>27.8</td>
<td>0.3</td>
</tr>
<tr>
<td>Iraq</td>
<td>43.6</td>
<td>0.7</td>
</tr>
<tr>
<td>Israel</td>
<td>20.4</td>
<td>13.6</td>
</tr>
<tr>
<td>Japan</td>
<td>10.7</td>
<td>27.3</td>
</tr>
<tr>
<td>Malaysia</td>
<td>28.0</td>
<td>2.5</td>
</tr>
<tr>
<td>Mexico</td>
<td>26.6</td>
<td>3.1</td>
</tr>
<tr>
<td>Nigeria</td>
<td>43.3</td>
<td>0.2</td>
</tr>
<tr>
<td>Pakistan</td>
<td>41.8</td>
<td>0.4</td>
</tr>
<tr>
<td>Philippines</td>
<td>30.4</td>
<td>0.7</td>
</tr>
<tr>
<td>Russia</td>
<td>12.6</td>
<td>8.6</td>
</tr>
<tr>
<td>South Africa</td>
<td>33.4</td>
<td>2.6</td>
</tr>
<tr>
<td>Spain</td>
<td>11.2</td>
<td>13.4</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>13.2</td>
<td>17.4</td>
</tr>
<tr>
<td>United States</td>
<td>15.2</td>
<td>22.6</td>
</tr>
</tbody>
</table>

The overall shape of the scatterplot looks roughly like that of exponential decay (exponential with a negative exponent), except for one problem: The curve should be close to the x-axis as it moves out into the right tail, a pattern that would correspond to birthrates dropping to zero as richer and richer countries are considered. That doesn't happen, even approximately, for the countries in the data set.

a. How might you first transform the data to give an exponential model a chance of fitting?

b. Make the transformation you identified in part a first, and then transform again to see if an exponential model is a reasonable one for these data.

c. What birthrate would you predict for a country with a gross national product of $15,000 per capita?

d. Does the same principle apply to the scatterplot in Display 11.74 of life expectancy at birth for men versus infant mortality rate for boys for the sample of countries in Display 11.41 on page 777?
Can you grow bigger flowers by reducing the length of their stems? Plant scientists designed an experiment to test different growth inhibitors to see which were most effective in reducing the length of stems, anticipating that reduced stem growth would enhance the quality of the flower.
Statistics today is big business. Nearly every large commercial enterprise and government agency in the United States needs employees who understand how to collect data, analyze it, and report conclusions. The language and techniques of survey and experimental design are part of politics, medicine, industry, advertising and marketing, and even, as you will see, flower growing. Consequently, a statistics course typically is required of college students who major in mathematics or in fields that use data, such as business, sociology, psychology, biology, and health science.

In this chapter, you will examine four examples of statistical practice in today’s world: case studies about how to

- produce bigger and better flowers
- gather information on Americans
- evaluate the economics of Major League Baseball
- study possible discrimination in employment

In each of these four real-world situations, statisticians have worked with the business or government agency involved. You will be following in their steps, so think of yourself as a consultant to the business or government agency. Compared to those in previous chapters, the data will be more complex and which statistical technique to use will be less clear-cut. Several statistical techniques may be appropriate, and you will have to decide which would be best. Often, the data don’t quite satisfy the conditions for the techniques you have learned, and you will have to think about what can be done in such situations, which come up in practice more often than not. If you do all four case studies, this chapter can serve as a comprehensive review of the statistical concepts and techniques that you have learned in this course.
The commercial flower industry is large and growing in the United States, particularly in Florida. The chrysanthemum, or mum, is a popular flower for bouquets and potted plants because it is hardy, long-lasting, and colorful. If you visit a flower shop to choose a bouquet of mums for someone special, you probably will look for the largest, most fully developed, most colorful blooms you can find. Flowers with these characteristics are in demand, and they don't show up in your flower shop by accident. Considerable research goes into producing fine blooms.

Plants can produce and use only so much energy. So within the same species, a plant with a long stem is likely to have a smaller flower than a plant with a short stem. The Environmental Horticulture Department of the University of Florida experimented with growth inhibitors to see which were most effective in reducing the length of stems, anticipating that reduced stem growth would enhance the quality of the flower. There were many treatments in the complete study: this investigation will consider only seven of them. The patterns you will see here remain much the same when all treatments are considered.

Individual plants of the same age were grown under nearly identical conditions, except for the growth-inhibitor treatment. Each treatment was randomly assigned to ten plants, whose heights (in centimeters) were measured at the outset of the experiment ($H_{t_1}$) and after a period of 10 weeks ($H_{t_2}$). (Heights actually were measured at intervening times as well, but those measurements are not part of this analysis.)

You can find the raw data for this experiment—the treatment each of the 70 plants received and the height of each plant before and after treatment—in Display 12.3 on page 802.

### DISCUSSION

#### Analyzing the Experiment

D1. Can you suggest possible improvements to this experimental design?

D2. What response variable would be best to use for comparing treatment groups in this experiment?
D3. What statistical method(s) would you suggest for comparing the effectiveness of, say, Treatments 1 and 2? What conditions must you check before proceeding?

Display 12.1 shows the statistical summaries of the growth, $H_{t_j} - H_{t_i}$, during the 10 weeks for each of the seven treatment groups. Notice that the means fluctuate quite a bit from group to group, as do the variances.

<table>
<thead>
<tr>
<th>Group</th>
<th>Count</th>
<th>Mean</th>
<th>Median</th>
<th>Variance</th>
<th>StdDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>10</td>
<td>30.8500</td>
<td>29.5000</td>
<td>58.4472</td>
<td>7.64508</td>
</tr>
<tr>
<td>02</td>
<td>10</td>
<td>35.9500</td>
<td>38</td>
<td>40.6917</td>
<td>6.37900</td>
</tr>
<tr>
<td>03</td>
<td>10</td>
<td>43.0500</td>
<td>44</td>
<td>25.4139</td>
<td>5.04122</td>
</tr>
<tr>
<td>04</td>
<td>10</td>
<td>52.7500</td>
<td>53.7500</td>
<td>18.2917</td>
<td>4.27688</td>
</tr>
<tr>
<td>05</td>
<td>10</td>
<td>28.4500</td>
<td>29.5000</td>
<td>58.3028</td>
<td>7.63563</td>
</tr>
<tr>
<td>06</td>
<td>10</td>
<td>35.9000</td>
<td>38.7500</td>
<td>58.7111</td>
<td>7.66232</td>
</tr>
<tr>
<td>07</td>
<td>10</td>
<td>39.7500</td>
<td>40.7500</td>
<td>20.9028</td>
<td>4.57196</td>
</tr>
</tbody>
</table>

Display 12.2 shows the boxplots of the differences in height, $H_{t_j} - H_{t_i}$, by treatment. The boxplots give a better view of the key features of the data than does the summary table.

Display 12.3 shows the raw data for this experiment.
<table>
<thead>
<tr>
<th>Treatment</th>
<th>Initial Height ($H_{ti}$)</th>
<th>Final Height ($H_{tf}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>48</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>32.5</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>20</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>32</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>37</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>29</td>
</tr>
<tr>
<td>1</td>
<td>2.5</td>
<td>34</td>
</tr>
<tr>
<td>1</td>
<td>4.5</td>
<td>46</td>
</tr>
<tr>
<td>1</td>
<td>3.5</td>
<td>33</td>
</tr>
<tr>
<td>1</td>
<td>3.5</td>
<td>30</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>33.5</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>35.5</td>
</tr>
<tr>
<td>2</td>
<td>3.5</td>
<td>47</td>
</tr>
<tr>
<td>2</td>
<td>2.5</td>
<td>41</td>
</tr>
<tr>
<td>2</td>
<td>4.5</td>
<td>42.5</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>24.5</td>
</tr>
<tr>
<td>2</td>
<td>4.5</td>
<td>47.5</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>42</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>37</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>40</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>47</td>
</tr>
<tr>
<td>3</td>
<td>2.5</td>
<td>44</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>48</td>
</tr>
<tr>
<td>3</td>
<td>2.5</td>
<td>46.5</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>44.5</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>54</td>
</tr>
<tr>
<td>3</td>
<td>3.5</td>
<td>48.5</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>46</td>
</tr>
<tr>
<td>3</td>
<td>3.5</td>
<td>47.5</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>32.5</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>55</td>
</tr>
<tr>
<td>4</td>
<td>2.5</td>
<td>57.5</td>
</tr>
<tr>
<td>4</td>
<td>1.5</td>
<td>58.5</td>
</tr>
<tr>
<td>4</td>
<td>2.5</td>
<td>60</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>56</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Initial Height ($H_{ti}$)</th>
<th>Final Height ($H_{tf}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2.5</td>
<td>48</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>54</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>56.5</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>57.5</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>32</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>37</td>
</tr>
<tr>
<td>5</td>
<td>1.5</td>
<td>15.5</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>41</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>28</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>31</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>23.5</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>35.5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>42.5</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>48</td>
</tr>
<tr>
<td>6</td>
<td>3.5</td>
<td>43.5</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>40.5</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>38</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>43</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>44</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>26.5</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>23</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>47</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>37</td>
</tr>
<tr>
<td>7</td>
<td>1.5</td>
<td>44.5</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>39</td>
</tr>
<tr>
<td>7</td>
<td>2.5</td>
<td>45.5</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>42.5</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>43</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>45</td>
</tr>
<tr>
<td>7</td>
<td>1.5</td>
<td>38</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>46</td>
</tr>
<tr>
<td>7</td>
<td>1.5</td>
<td>38.5</td>
</tr>
<tr>
<td>7</td>
<td>2.5</td>
<td>32</td>
</tr>
</tbody>
</table>

Display 12.3  The treatment each of the 70 plants received and its height before and after treatment. [Source: University of Florida Institute for Food and Agricultural Sciences, 1997.]
Practice

Analyzing the Experiment

P1. Which two treatments appear to be most effective in reducing the stem length? Least effective? Explain your choices.

P2. Compare the distributions in terms of center and spread. Do you see any pattern relating the spreads to the centers? Make a statistical graphic that shows how the mean and standard deviation are related. Describe this relationship. Is this the kind of relationship you would have expected? Explain.

P3. Refer to P2. If one treatment has a mean that is 1 cm larger than that of another treatment, how would you expect the standard deviations to differ?

P4. Is the difference in mean growth between the two most effective treatments statistically significant? Be sure to check the conditions for the test you select. Also write the hypotheses in words.

P5. The two treatments that differ most appear to be Treatments 4 and 5. Why is it clear from the plot that the difference in mean growth between these two treatments is statistically significant? Construct a 95% confidence interval to support your reasoning.

P6. By hand, make boxplots of the initial heights of the mums assigned to Treatments 4 and 5. Show any outliers determined by the 1.5 \times IQR rule. Describe any differences in the initial heights of the plants assigned to these two treatments. Is there enough difference to call into question the conclusion in P5?

12.2 Keeping Tabs on Americans

The American Community Survey (ACS) is a nationwide sample survey that collects fairly detailed data on persons and households at nearly 3 million addresses across the United States. The goals of the annual survey are to provide federal, state, and local governments with an information base for the administration and evaluation of government programs; to facilitate improvement of census data; and to provide timely demographic, housing, social, and economic statistics that can be compared across states, communities, and population groups.

To keep things manageable, in this case study you will work with a sample of 40 households and with only 12 of the more than 50 variables measured by the ACS. This sample of 40 households comes from Florida (where over 27,000 households are sampled in the complete ACS). The 40 households are a random sample of a random sample, and so you can think of them as a simple random sample of Florida households. The data are provided in Display 12.4 (on the next page) for these variables:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP</td>
<td>number of persons</td>
</tr>
<tr>
<td>BDS</td>
<td>number of bedrooms</td>
</tr>
<tr>
<td>BLD</td>
<td>building type: (1 = mobile home, 2 = single-family detached house, 3 = apartment or other attached structure)</td>
</tr>
<tr>
<td>ELEP</td>
<td>mean monthly electricity payment</td>
</tr>
<tr>
<td>INSP</td>
<td>hazard insurance payment for the year</td>
</tr>
<tr>
<td>MRGP</td>
<td>monthly mortgage payment</td>
</tr>
<tr>
<td>RMS</td>
<td>number of rooms</td>
</tr>
</tbody>
</table>
OWN: ownership (1 = own with a mortgage, 2 = own free and clear, 3 = rent)

VEH: number of motor vehicles

HHL: household primary language (1 = English, 2 = other)

NOC: number of children under the age of 18

R65: at least one household age 65 or older (0 = no, 1 = yes)

<table>
<thead>
<tr>
<th>NP</th>
<th>BDS</th>
<th>BLD</th>
<th>ELEP</th>
<th>INS</th>
<th>MRGP</th>
<th>RMS</th>
<th>OWN</th>
<th>VEH</th>
<th>HHL</th>
<th>NOC</th>
<th>R65</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5</td>
<td>2</td>
<td>120</td>
<td>1400</td>
<td>8</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3</td>
<td>100</td>
<td>1700</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>90</td>
<td>200</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>90</td>
<td>190</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>50</td>
<td>2800</td>
<td>7</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>90</td>
<td>1700</td>
<td>7</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Random sample of 40 ACS households in Florida.

Analyzing the Survey Data

D4. Give at least two possible reasons for the missing data in the table in Display 12.4. Do you suppose such gaps are common in surveys of this type?

D5. Based on these data and the techniques of this book, can you construct a confidence interval for the difference between the proportion of households residing in detached houses and the proportion residing in mobile homes? Explain why or why not.

D6. Can you construct a confidence interval for the difference between the mean hazard insurance payments for owners and renters with these data? Explain why or why not.

D7. In addition to retired folks on limited incomes, Florida has some very wealthy residents. Why, then, is the largest number of rooms among households in this data set only nine?

Inference with Univariate Categorical Data

Florida has the reputation of being a state full of elderly retired folks, so let’s begin by estimating the proportion of households that actually contain at least one person age 65 or older. The sample proportion, \( \hat{p} \), is 13/40, or 0.325. The conditions for estimation are met because this is a random sample and both \( n\hat{p} = 13 \) and \( n(1 - \hat{p}) = 27 \) are greater than 10. The 95% confidence interval estimate turns out to be (0.18, 0.47), a fairly wide interval because of the relatively small sample size. So it is plausible that nearly half of Florida households contain a person age 65 or older.

Inference with Bivariate Categorical Data

Many of the elderly live on nearly fixed incomes, and mobile homes are relatively inexpensive compared to houses. Because the retired tend to have less income, perhaps a higher proportion of them live in mobile homes. To check this out, you can look for a possible association between \( R65 \) and \( BLD \) (building type) by summarizing the data in a two-way table. Display 12.5 (on the next page) gives such a table, with the expected cell frequencies under the null hypothesis of no association and the resulting chi-square test of independence.

The \( P \)-value is small, but not small enough to reject the null hypothesis of independence between the categories at the 0.05 significance level. In fact, this \( \chi^2 \) statistic is inflated a bit because three of the cells have expected frequencies less than 5, so the actual \( P \)-value might be even greater than the one shown. In short, there is little evidence to support an association between whether a household has an older member and the type of buildings in which the household resides.

As in Chapter 1, you can use randomization as a basis for making a decision about two-way tables. This method does not depend on the \( \chi^2 \) distribution and, in particular, does not require expected cell frequencies of 5 or greater. If there is no association between \( BLD \) and \( R65 \), then the observed value of the \( \chi^2 \) statistic (5.432 in this example) is simply a random occurrence and does not depend
Test of Sample of ACS

First attribute (categorical): BLD
Second attribute (categorical): R65

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Row Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLD</td>
<td>6</td>
<td>16</td>
<td>5</td>
<td>27</td>
</tr>
<tr>
<td>R65</td>
<td>0</td>
<td>1</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>21</td>
<td>12</td>
<td>40</td>
</tr>
</tbody>
</table>

First attribute: BLD
Number of categories: 3
Second attribute: R65
Number of categories: 2

Warning: 3 out of 6 cells have expected values less than 5.

Ho: BLD is independent of R65
Chi-square: 5.432
DF: 2
P-value: 0.066

The numbers in parentheses in the table are expected counts.

Display 12.5 Summary of households with members at least age 65 versus building type.

on the way the pairs of data on BLD and R65 were matched up in the observed sample. To check the possibility that a $\chi^2$ statistic this large reasonably could occur by chance, you must generate a randomization distribution of possible $\chi^2$ values. Begin by listing the BLD values in a column. Write the R65 values on cards and shuffle them. Deal them out, one by one, so that you pair each BLD value with a randomly assigned R65 value. Then compute $\chi^2$ for this randomization.

You do this many times. Display 12.6 shows 200 values for such a randomization distribution.

Display 12.6 Randomized values of $\chi^2$ for BLD versus R65.
The observed value, 5.432, was met or exceeded 15 times out of 200 runs, making the approximate $P$-value 0.075. Notice that this is larger than the $P$-value from the formal chi-square test (0.066) and provides even less evidence against the null hypothesis of no association between whether the household has a member age 65 or older and the type of building in which the household resides.

**DISCUSSION**

**Inference with Bivariate Categorical Data**

D8. Create one additional value for the randomization distribution in Display 12.6.

D9. Florida sometimes conjures up the image of many retired people living in mobile homes. How would you proceed if you wanted to determine if householders age 65 and older are more likely to live in mobile homes than not?

**Inference with Univariate Measurement Data**

How many rooms, on average, does a household in Florida have for its use? A plot of the sample data on RMS, given in Display 12.7, shows that the sample might well have come from a normal distribution.

![Display 12.7 Number of rooms in the sample of households from the American Community Survey.](image)

Conditions for a confidence interval estimate of plausible values for the population mean are met, and the 95% confidence interval for the number of rooms per household in Florida is (4.76, 5.79). Even though there are some very large houses in Florida, the preponderance of smaller places of residence keeps the mean number of rooms per household rather small.

**Inference with Bivariate Measurement Data**

Most households in Florida use a lot of air conditioning, so electricity costs can be considerable. It seems reasonable that the electricity bill should increase with the number of rooms, but will the pattern of association be linear?

Display 12.8 (on the next page) shows a fairly pronounced linear trend for these data. The regression equation is $ELEP = 17.1 \cdot RMS + 34.7$. You will work with this regression in P18 and P19.
Display 12.8 Electricity payment versus number of rooms for the ACS sample.

Practice

Inference with Univariate Categorical Data

P7. Suppose someone claims that a quarter of Florida households have a primary language other than English. Can you refute this claim based on the data in Display 12.4 on page 804?

P8. Estimate with 95% confidence the proportion of Florida households residing in apartments or other attached structures.

P9. What plausible proportions of Florida households own their home outright?

Inference with Bivariate Categorical Data

P10. Construct a two-way table of BLD versus OWN. By looking carefully at your two-way table but not actually conducting a test, does it appear that these two variables are associated? If so, explain the nature of the association.

P11. In some cultures it is common to have extended families living in the same household, which might suggest that a greater proportion of non–English speaking households would contain an older resident. Do the data in Display 12.4 contain any evidence of this? Which two equivalent significance tests could you use to answer this question? What concerns do you have, if any, about conducting and interpreting these tests?

P12. Is there a significant association between primary language spoken in the household (HHL) and the type of building in which the household is found (BLD)? If so, explain the nature of the apparent association. Are you concerned about the accuracy of the reported P-value here? Why or why not?

P13. Display 12.9 (on the next page) shows 200 runs of the randomization distribution for $\chi^2$ values with BLD randomly paired with HHL.

a. Describe how this distribution was constructed.

b. Does this distribution alleviate any concerns you might have about the results of P12? Explain.
Inference with Univariate Measurement Data

P14. Estimate the mean number of people per household in Florida, with 90% confidence. Be sure to check the conditions first.

P15. Focus on the monthly mortgage payments, for those households that have them.
   a. Plot the data on monthly mortgage payments and describe the distribution. Do these data appear to meet the conditions for inference for a mean? Will a transformation help?
   b. In 2003, the mean monthly mortgage payment nationwide was $840. Is there statistically significant evidence that the mean was less than that in Florida?

P16. Florida is a hot spot for hurricanes, so hazard insurance for a person’s place of residence is highly recommended. Consider the yearly amounts paid for hazard insurance in the sampled households in Display 12.4.
   a. Plot the data on insurance payments and describe the distribution. Is a transformation needed here, in order to allow you to make inferences about the mean? If so, find a transformation that works well.
   b. Estimate the population mean of the appropriately transformed data in a 95% confidence interval.

P17. Estimate the difference between the mean number of rooms in detached houses and in apartments in a 95% confidence interval.

Inference with Bivariate Measurement Data

P18. Interpret the slope of the regression line in Display 12.8.

P19. Is the slope in Display 12.8 significantly different from 0? Be sure to do all steps of the significance test.

P20. Make an appropriate plot and study the relationship between the number of people and the number of bedrooms in the households in Display 12.4 on page 804. Would you say that the number of bedrooms in a residence is a strong predictor of the number of people in a household? Explain your reasoning.

P21. One possible use of the data in Display 12.4 is to build a model to predict the number of vehicles that might be associated with a household.
   a. Which is the better predictor of the number of vehicles attached to a household: the number of people, or the number of rooms?
   b. Test for the significance of the slope in each relationship in part a. What are your conclusions? Give a possible practical reason for this state of affairs.
P22. How does the yearly hazard insurance payment relate to the number of rooms?

a. Plot INSP versus RMS, fit a line to the data, and interpret the slope.

b. Do the conditions for a confidence interval for the slope and for a test of significance of the slope appear to be satisfied here? If not, explain what is amiss.

c. Find a transformation that “cures” whatever is amiss in part b. Then redo the regression and estimate the slope of the line relating the two variables in a 95% confidence interval and interpret the result.

12.3 Baseball: Does Money Buy Success?

Salaries of professional athletes are astronomical, but huge payrolls are not necessarily bad for team owners if the salaries are associated with great team performance or with other variables that improve income, such as attendance. What are the relationships between some of these key variables for Major League Baseball?

Display 12.10 gives the total payroll, the average salary per player, the total attendance, the team batting average (percentage of hits, without the decimal), and the percentage of games won in regular-season play (without the decimal) of teams for the 2001 season. Teams are divided into two major leagues (the American and the National), and each league is divided into three divisions (East, Central, and West).

Exploring the Table of Data

D10. Looking only at Display 12.10, do you see any interesting features of the data? Do you see any possible patterns or associations between variables?

D11. Using statistical software, explore these data and comment on any interesting patterns or possible relationships you find. Are there any surprises?

D12. Which variable was most strongly associated with payroll? Does it matter much whether you use total team payroll or average salary per player in looking for relationships with other variables in the data set? Explain.

D13. Discuss whether you think it is appropriate to do inferential analyses on these data.
## Display 12.10


*Sources: www.CNNSI.com; www.slam.ca; www.mlb.com, June 20, 2002.*

<table>
<thead>
<tr>
<th>Team</th>
<th>Payroll ($ millions)</th>
<th>Average Salary ($)</th>
<th>Attendance (thousands)</th>
<th>Batting Average</th>
<th>% Wins (in tenths of a percent)</th>
<th>League</th>
<th>Division</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York (AL)</td>
<td>109.79</td>
<td>3,541,674</td>
<td>3167</td>
<td>267</td>
<td>594</td>
<td>A</td>
<td>E</td>
</tr>
<tr>
<td>Boston</td>
<td>109.55</td>
<td>3,423,716</td>
<td>2592</td>
<td>266</td>
<td>509</td>
<td>A</td>
<td>E</td>
</tr>
<tr>
<td>Los Angeles</td>
<td>108.98</td>
<td>3,757,964</td>
<td>3017</td>
<td>255</td>
<td>531</td>
<td>N</td>
<td>W</td>
</tr>
<tr>
<td>New York (NL)</td>
<td>93.17</td>
<td>3,327,658</td>
<td>2618</td>
<td>249</td>
<td>506</td>
<td>N</td>
<td>E</td>
</tr>
<tr>
<td>Cleveland</td>
<td>91.97</td>
<td>3,065,833</td>
<td>3182</td>
<td>278</td>
<td>562</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>Atlanta</td>
<td>91.85</td>
<td>2,962,958</td>
<td>2779</td>
<td>260</td>
<td>543</td>
<td>N</td>
<td>E</td>
</tr>
<tr>
<td>Texas</td>
<td>88.50</td>
<td>2,854,981</td>
<td>2831</td>
<td>275</td>
<td>451</td>
<td>A</td>
<td>W</td>
</tr>
<tr>
<td>Arizona</td>
<td>81.20</td>
<td>2,900,233</td>
<td>2736</td>
<td>267</td>
<td>568</td>
<td>N</td>
<td>W</td>
</tr>
<tr>
<td>St. Louis</td>
<td>77.27</td>
<td>2,664,512</td>
<td>3110</td>
<td>270</td>
<td>574</td>
<td>N</td>
<td>C</td>
</tr>
<tr>
<td>Toronto</td>
<td>75.79</td>
<td>2,707,089</td>
<td>1915</td>
<td>263</td>
<td>494</td>
<td>A</td>
<td>E</td>
</tr>
<tr>
<td>Seattle</td>
<td>75.65</td>
<td>2,701,875</td>
<td>3512</td>
<td>288</td>
<td>716</td>
<td>A</td>
<td>W</td>
</tr>
<tr>
<td>Baltimore</td>
<td>72.42</td>
<td>2,497,460</td>
<td>3064</td>
<td>248</td>
<td>391</td>
<td>A</td>
<td>E</td>
</tr>
<tr>
<td>Colorado</td>
<td>71.06</td>
<td>2,632,148</td>
<td>3168</td>
<td>292</td>
<td>451</td>
<td>N</td>
<td>W</td>
</tr>
<tr>
<td>Chicago (NL)</td>
<td>64.01</td>
<td>2,462,147</td>
<td>2779</td>
<td>261</td>
<td>543</td>
<td>N</td>
<td>C</td>
</tr>
<tr>
<td>San Francisco</td>
<td>63.33</td>
<td>2,345,654</td>
<td>3269</td>
<td>266</td>
<td>556</td>
<td>N</td>
<td>W</td>
</tr>
<tr>
<td>Chicago (AL)</td>
<td>62.36</td>
<td>2,309,741</td>
<td>1766</td>
<td>268</td>
<td>512</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>Houston</td>
<td>60.38</td>
<td>2,236,395</td>
<td>2906</td>
<td>271</td>
<td>574</td>
<td>N</td>
<td>C</td>
</tr>
<tr>
<td>Tampa Bay</td>
<td>54.95</td>
<td>2,035,245</td>
<td>1298</td>
<td>258</td>
<td>383</td>
<td>A</td>
<td>E</td>
</tr>
<tr>
<td>Pittsburgh</td>
<td>52.69</td>
<td>1,699,946</td>
<td>2436</td>
<td>247</td>
<td>383</td>
<td>N</td>
<td>C</td>
</tr>
<tr>
<td>Detroit</td>
<td>49.83</td>
<td>1,779,685</td>
<td>1921</td>
<td>260</td>
<td>407</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>Anaheim</td>
<td>46.56</td>
<td>1,502,199</td>
<td>1998</td>
<td>261</td>
<td>463</td>
<td>A</td>
<td>W</td>
</tr>
<tr>
<td>Cincinnati</td>
<td>45.22</td>
<td>1,739,534</td>
<td>1880</td>
<td>262</td>
<td>407</td>
<td>N</td>
<td>C</td>
</tr>
<tr>
<td>Milwaukee</td>
<td>43.08</td>
<td>1,595,901</td>
<td>2811</td>
<td>251</td>
<td>420</td>
<td>N</td>
<td>C</td>
</tr>
<tr>
<td>Philadelphia</td>
<td>41.66</td>
<td>1,602,468</td>
<td>1792</td>
<td>260</td>
<td>531</td>
<td>N</td>
<td>E</td>
</tr>
<tr>
<td>San Diego</td>
<td>38.33</td>
<td>1,419,745</td>
<td>2378</td>
<td>252</td>
<td>488</td>
<td>N</td>
<td>W</td>
</tr>
<tr>
<td>Kansas City</td>
<td>35.64</td>
<td>1,229,069</td>
<td>1536</td>
<td>266</td>
<td>401</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>Florida</td>
<td>35.50</td>
<td>1,183,472</td>
<td>1261</td>
<td>264</td>
<td>469</td>
<td>N</td>
<td>E</td>
</tr>
<tr>
<td>Montreal</td>
<td>34.77</td>
<td>1,159,150</td>
<td>643</td>
<td>253</td>
<td>420</td>
<td>N</td>
<td>E</td>
</tr>
<tr>
<td>Oakland</td>
<td>33.81</td>
<td>1,252,250</td>
<td>2133</td>
<td>264</td>
<td>630</td>
<td>A</td>
<td>W</td>
</tr>
<tr>
<td>Minnesota</td>
<td>24.35</td>
<td>901,852</td>
<td>1783</td>
<td>272</td>
<td>525</td>
<td>A</td>
<td>C</td>
</tr>
</tbody>
</table>
Is Inference Appropriate?

This is an observational study because there was no random sampling and no random assignment of treatments. (Owners probably would object to a random assignment of a payroll to each team.) The analysis should start, as you already have done, with the exploration of possible relationships among the variables, but some inference might be possible if the nature of the inference is carefully phrased. Payroll is certainly not random; for the most part, these values are fixed at the beginning of the season. Even though attendance, batting average, and wins are somewhat random events that occur throughout the season, they were neither randomly sampled nor randomly assigned. You can, however, still ask, “How likely am I to observe by mere chance a pattern like the one seen in the data?” In studying bivariate relationships, the null hypothesis in the significance test is that the data look like a random sample of y’s for fixed values of x from a population with no linear trend. You can test to see whether this is a reasonable model or whether a linear trend appears to be a better explanation. In comparing two means, the null hypothesis is that the data look like independent random samples from two populations with the same mean.

This case study is similar in construction to common investigations of industrial processes in which yields are measured after temperature and pressure gauges have been set at fixed levels determined by an engineer. Often, these temperature settings cannot be randomized because of other considerations in the process, such as the time it takes to get a process up to the required temperature and pressure levels. Even with no randomness in the design, the statistical analysis of yields does help decide whether an observed association can be attributed to chance alone. But it cannot answer questions of cause and effect.

Differences Between the Leagues

Of interest in this study is whether possible associations among the variables are different for the two leagues. Display 12.11 shows a scatterplot, residual plot, and regression analysis of percent wins versus payroll for all 30 teams.

Display 12.11  Percent wins versus payroll for all 30 teams.
The observed trend is increasing, but a closer scrutiny of the test of significance for the slope shows that the null hypothesis of a zero slope has a two-sided $P$-value of 0.0674. This is small, but not small enough to make a bold declaration in favor of a linear trend with nonzero slope. In other words, the trend here is no more pronounced than you would expect if you randomly assigned the winning percentages to the different teams while keeping their payroll fixed.

Separating the leagues, however, gives the plots and analyses in Display 12.12. While the American League shows no significant linear trend, the National League does show a significant trend. That is, if payroll were kept constant for each team in the American League but percentage of wins were reassigned at random to the teams, then it is quite likely that you would get a slope as far from 0 as the American League’s 0.782. If you did the same thing for the teams in the National League, it isn’t likely that you would get a slope as far from 0 as 1.49, indicating that the positive trend in the National League cannot reasonably be explained by chance.

Display 12.12 Percent wins versus payroll for the American League and the National League.
**DIFFERENCES BETWEEN THE LEAGUES**

D14. Compare the plots in Display 12.12 for the two leagues. What is it about these data that produces the drastic difference in results for the two leagues?

D15. Besides having a random sample, what other conditions need to be met for a regression analysis? From the plots in Display 12.12, does it seem reasonable to assume that these other conditions are met?

**SIMULATING A P-VALUE**

The null hypothesis for percent wins versus payroll is this:

The observed slope is no farther from 0 than you would be reasonably likely to get if you randomly reassigned the values of percent wins to different teams while keeping each team’s payroll fixed.

The $P$-value for the test of significance for the slope measures how unusual the observed slope would be under those conditions. You can estimate the $P$-value by repeatedly rearranging the $y$-values and observing what happens to the slopes. Using this idea, Display 12.13 shows a set of 100 slopes found by 100 rerandomizations of the values of percent wins for the American League.

<table>
<thead>
<tr>
<th>Stem-and-leaf of slopes for American League</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leaf Unit = 0.10</td>
</tr>
<tr>
<td>1  -2  2</td>
</tr>
<tr>
<td>7  -1  976655</td>
</tr>
<tr>
<td>18 -1  4443211000</td>
</tr>
<tr>
<td>31 -0  9988877666555</td>
</tr>
<tr>
<td>(22)  -0  433333333332211110000</td>
</tr>
<tr>
<td>47  0  00000112233333444</td>
</tr>
<tr>
<td>29  0  55566677888899</td>
</tr>
<tr>
<td>15  1  001122444</td>
</tr>
<tr>
<td>5  1  779</td>
</tr>
<tr>
<td>2  2  01</td>
</tr>
</tbody>
</table>

**Display 12.13** Stem-and-leaf plot of 100 simulated slopes for percent wins versus payroll for the American League.

In the 100 trials, the observed slope, 0.782, for the American League was equaled or exceeded 21 times, giving a simulated one-sided $P$-value of 0.21. The $t$-test in Display 12.12 gives a two-sided $P$-value of 0.4545, which is almost equal to the $P$-value from the simulation. So with either method, the question of whether the observed pattern reasonably could be attributed to mere chance is answered in the affirmative for the American League.

**DISCUSSION**

D16. The stem-and-leaf plot in Display 12.14 shows a simulation for the National League that is parallel to the simulation for the American League.

a. Describe how the simulation was conducted.

b. Conduct one more trial and show where it would go on the stemplot.
c. What is the simulated one-sided $P$-value for these 101 trials? How does it compare to the two-sided $P$-value given in Display 12.12? What is your conclusion about the observed pattern for the National League?

Stem-and-leaf of slopes for National League
Leaf Unit = 0.10
2 -1 44
2 -1
3 -1 1
8 -0 99988
14 -0 76666
18 -0 4444
31 -0 333333332222
45 -0 1111111100000
(13) 0 0000000000011
42 0 2222222233
32 0 44445555
24 0 677777
18 0 888999
12 1 00000111
4 1 3
3 1 4
2 1
2 1 89

Display 12.14 One hundred simulated slopes for percent wins versus payroll for the National League.

**Ecological Correlations**

Each of the two major leagues is divided into three divisions, so another way of looking at the baseball data is to explore what happens at the division level. Display 12.15 shows the averages for the data in Display 12.10 on page 811 by division. (All teams play a 162-game schedule, with similar numbers at bat for the season, so it is fair to take simple averages of team batting averages and winning percentages.)

<table>
<thead>
<tr>
<th>Division</th>
<th>Average Payroll ($ millions)</th>
<th>Average Attendance (in thousands)</th>
<th>Batting Average</th>
<th>% Wins (in tenths of a percent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC</td>
<td>52.8</td>
<td>2037.6</td>
<td>269</td>
<td>481</td>
</tr>
<tr>
<td>AE</td>
<td>84.5</td>
<td>2407.2</td>
<td>260</td>
<td>474</td>
</tr>
<tr>
<td>AW</td>
<td>61.1</td>
<td>2618.5</td>
<td>272</td>
<td>565</td>
</tr>
<tr>
<td>NC</td>
<td>57.1</td>
<td>2653.7</td>
<td>260</td>
<td>484</td>
</tr>
<tr>
<td>NE</td>
<td>59.4</td>
<td>1818.6</td>
<td>257</td>
<td>494</td>
</tr>
<tr>
<td>NW</td>
<td>72.6</td>
<td>2913.6</td>
<td>266</td>
<td>519</td>
</tr>
</tbody>
</table>


In analyzing how percent wins is related to batting average, you could use the teams as cases or use the divisions as cases. The scatterplot for each analysis is shown in Display 12.16.
Display 12.16  Scatterplots of percent wins versus batting average, by team and by division.

Each plot in Display 12.16 shows an increasing trend, but the correlation in the first is about 0.49, while the correlation in the second is about 0.66. In this situation, using the division averages instead of the individual team data inflates the correlation and makes the linear trend appear to be stronger.

The correlation computed using the divisions as cases and the division averages as variables is an example of an ecological correlation, or a correlation between group averages. Such correlations are often used in subjects such as sociology and political science, where it is easier to get information about groups of people than about individuals. (For example, it’s easy to find the percentage of people in your state who voted Republican in the last presidential election but almost impossible to get the same information for each individual voter.) As you have seen, correlations based on groups and averages can be quite different from correlations based on individuals.

Ecological Fallacy

Ecological correlations use groups as cases and averages as variables. For many situations, you get quite different values than you would if you used individuals as cases. The mistake of using ecological correlations to support conclusions about individuals is called the ecological fallacy.

Of course, the team statistics are themselves averages (or totals), and we could have analyzed the relationship between salaries and batting averages by using individual players as cases. For the purpose of studying teams as business entities—taking into account variables such as attendance and percent wins—it makes sense to use teams as cases. Team statistics are of little use, however, in studying player performance. So the choice of what to use as cases depends on the objectives of the study.

Ecological Correlations

D17. Do both lines in Display 12.16 have slopes that are significantly greater than 0? Does this suggest another problem with using ecological correlations?
D18. Calculate the correlation between *batting average* and *payroll* using teams as cases and then using divisions as cases. Comment on possible reasons for any difference.

D19. Calculate the correlation between *percent wins* and *payroll* using teams as cases and then using divisions as cases. Does the latter correlation appear to have a meaningful interpretation?

### Practice

#### Differences Between the Leagues

For P23–P25: Analyze the given relationship

- a. for all teams
- b. for the American League
- c. for the National League
- d. Compare the three relationships.

**P23.** *payroll* versus *attendance*

**P24.** *batting average* versus *payroll*

**P25.** *percent wins* versus *batting average*

**P26.** Use an appropriate test to answer these questions about means.

- a. Is the difference between mean attendance for the two leagues statistically significant?
- b. Is the difference between mean batting average for the two leagues statistically significant?

**P27.** In the 2001 World Series, the New York Yankees lost to the Arizona Diamondbacks. Is the difference in their percentage of wins during the regular 162-game 2001 season statistically significant?

**P28.** For the 2001 regular season of 162 games, the three National League division winners were Atlanta, St. Louis, and Arizona. For the American League, the division winners were New York, Cleveland, and Seattle.

- a. Are there statistically significant differences among the proportions of games won by the three National League division winners?
- b. Are there statistically significant differences among the proportions of games won by the three American League division winners?

### 12.4 Martin v. Westvaco Revisited: Testing for Discrimination Against Employees

In Chapter 1, you read about Robert Martin, who was laid off from his job at the Westvaco Corporation at age 54. The statistical analysis presented during the lawsuit was quite a bit more involved than the simplified version you saw in Chapter 1. With what you have learned, you can now carry out a much more thorough analysis while reviewing some inferential techniques as well as some ideas from probability theory.

In the *Westvaco* case, Martin claimed that he had been terminated because of his age, which, if true, would be against the law. His case eventually was settled out of court, before going to trial, but not until after a lot of statistical analysis and some arguments about the statistics.

Display 1.1 on page 5 shows the data used in the lawsuit, arranged in a standard “cases by variables” format. Each row of the table represents a case—one of the 50 people who worked in the engineering department of the envelope division of Westvaco when layoffs began.
Comparing Termination Rates for Two Age Groups

By law, all employees age 40 or older belong to what is called a “protected class”. To discriminate against them on the basis of their age is against the law. At the time of the layoffs at Westvaco, 36 of the 50 people working in the engineering department were age 40 or older. A total of 28 workers were terminated, and 21 of them were age 40 or older.

Comparing Termination Rates for Two Age Groups

D20. Construct a table to satisfy each of these descriptions. In each table, keep the marginal totals the same as in the Westvaco case, presented in Display 12.17.

a. Make a copy of Display 12.17 and fill in the cells in order to present the strongest possible case for discrimination against workers age 40 or older.

<table>
<thead>
<tr>
<th>Age</th>
<th>Terminated?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under 40</td>
<td>Yes</td>
</tr>
<tr>
<td>Under 40</td>
<td>No</td>
</tr>
<tr>
<td>Total</td>
<td></td>
</tr>
</tbody>
</table>

Display 12.17 Two-way table of layoff status and age, with marginal totals from the Westvaco case.

b. Make another table, but this time fill in the cells in a way that makes the variables terminated and age as close to independent as possible. If you select an employee at random from those represented in your table, are the variables terminated and age independent events according to the definition in Section 5.5?

c. Make a final table using the actual data from Westvaco. Does it look more like the table in part a or the table in part b? Do you have strong evidence that older workers were more likely to be chosen for layoff?

D21. Consider which statistical test would justify your opinion in D20, part c, about whether it looks as if older workers were more likely to be chosen for layoff.

a. Which tests of significance are possibilities in this situation? Give the strengths and weaknesses of each test as applied to this scenario. (Don’t worry about conditions for now.) Should you use a one-sided test or a two-sided test?

b. Using the data from your table in D20, part c, find the P-value for each of the tests in part a and compare them.

c. Select the test you think is most appropriate. What is your conclusion if you take the test at face value (that is, if you don’t worry about the conditions being met)?
Conditions Rarely Match Reality

The major differences between reality and a model are often the basis of heated arguments between opposing lawyers in discrimination cases. One test you might have chosen in D21 is the $z$-test for the difference between two proportions. This test is based on several assumptions.

**Independent Random Samples**
A simple random sample of size $n_1$ is taken from a large population with proportion of successes $p_1$. A second, independent random sample of size $n_2$ is taken from a large population with proportion of successes $p_2$.

**Large Samples from Even Larger Populations**
In the $z$-test, you use the normal distribution to approximate the sampling distribution of $\hat{p}_1 - \hat{p}_2$, where $\hat{p}_1$ is the proportion of successes in the first sample and $\hat{p}_2$ is the proportion of successes in the second sample. This approximation is reasonable provided that all of $n_1 \hat{p}_1$, $n_1(1 - \hat{p}_1)$, $n_2 \hat{p}_2$, and $n_2(1 - \hat{p}_2)$ are at least 5 and that each population is at least 10 times the sample size. On the surface, the mismatch between the two sets of conditions is striking:

$$
\begin{array}{|c|c|}
\hline
\text{z-Test for a Difference} & \text{Westvaco Case} \\
\text{Between Two Proportions} & \\
\hline
\text{Two populations that are} & \text{One population that is} \\
\bullet \text{both very large} & \bullet \text{small} \\
\hline
\text{Two samples that are} & \text{Two samples (under age 40 or 40 or older) that are} \\
\bullet \text{large enough} & \bullet \text{large enough} \\
\bullet \text{randomly selected} & \bullet \text{not randomly selected} \\
\bullet \text{from the two different populations} & \bullet \text{from the same population} \\
\bullet \text{independent of each other} & \bullet \text{as dependent as can be, because if you know} \\
& \text{one sample, you automatically know the other} \\
\hline
\end{array}
$$

This mismatch does not make the test invalid, however. You can still use the test to answer this question: “If the process had been random, how likely would it have been to get a difference in proportions as big as the one Westvaco got just by chance?” As long as it is made clear that this is the question being answered, the test is valid and can be very informative.

So you can proceed with a significance test, but you must make the limitations of what you are doing very clear. If you reject the null hypothesis, all you can conclude is that something happened that can’t reasonably be attributed to chance alone.

Another alternative is to use Fisher’s exact test, in which the sampling distribution of the difference of two proportions is constructed exactly rather than relying on a normal approximation. Fisher’s exact test requires few assumptions and uses no approximations. In contrast, the $z$-test for the difference between two proportions is an approximation and requires strong assumptions. So why don’t we always use Fisher’s exact test? We can, but it requires some computing power. You saw a simulation of this test in E19 in Chapter 1. As technology becomes more powerful, statisticians increasingly are turning to methods such as Fisher’s exact test rather than using approximations based on the normal distribution.
Conditions Rarely Match Reality

D22. Evaluate the conditions necessary for your significance test in D21, part c. Give a careful statement of the conclusion you can make.

D23. Agree or disagree, and tell why: “A hypothesis test is based on a probability model. Like all probability models, it assumes certain outcomes are random. But in the Westvaco case, the decisions about which people to lay off weren’t random. There’s no probability model, so a statistical test is invalid.”

Looking for a Better Approach

The test you did in D21 was based on dividing workers into two age groups, “under 40” and “40 or older.” This replaces a quantitative variable (age) with a categorical variable (age group). The cutoff age that defines who is classified as an “older” worker can be arbitrary and, as you’ll see in P30–P35, can change the results of the analysis. Looking at the mean ages of those laid off and those retained can help you avoid making such an arbitrary decision. You took that approach in Chapter 1 when you used the average age of the workers laid off by Westvaco as the test statistic. You’ll explore this approach further in P36–P38.

Conditions for a t-Test

A t-test for the difference between two means is based on independent random samples from two large populations, which must be approximately normal if the sample sizes are small. If you were to test that the difference between the mean ages of those laid off and those retained is 0, strict application of the t-test would require the conditions that those workers terminated and those retained are independent random samples from two large, approximately normal, populations. In actual fact, there is only one population that is finite rather than infinite, and this population is sorted into two groups. Again, the mismatch between the two sets of conditions is striking:

<table>
<thead>
<tr>
<th>t-Test for a Difference Between Means</th>
<th>Westvaco Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two populations that are</td>
<td>One population that is</td>
</tr>
<tr>
<td>• both very large</td>
<td>• small</td>
</tr>
<tr>
<td>• both normally distributed</td>
<td>• not normally distributed</td>
</tr>
<tr>
<td>Two samples that are</td>
<td>Two samples that are</td>
</tr>
<tr>
<td>• randomly selected</td>
<td>• not randomly selected</td>
</tr>
<tr>
<td>• from two different populations</td>
<td>• from the same population</td>
</tr>
<tr>
<td>• independent of each other</td>
<td>• as dependent as can be, because if you know one sample—those laid off—you automatically know the other</td>
</tr>
</tbody>
</table>

On the surface, there’s no apparent reason to think the t-test is appropriate for answering the question “Can the observed difference in means be attributed simply to chance, or should we look for another explanation?” Remarkably, though, extensive simulations have shown that despite the mismatch between the two sets of conditions, the t-test tends to give a good approximation of the P-value.
you would get using an approach that satisfies the actual set of conditions for situations like the Westvaco case.

You used the simulation approach in Section 1.2 when you selected three workers for layoff using a completely random process and computed their average age. After you did this step many times, you compared the average age of workers actually laid off at Westvaco to your distribution of the average ages of workers selected at random for layoff. The null hypothesis was that the process used by Westvaco was equivalent to randomly choosing the employees for layoff from the ten hourly workers remaining at the beginning of Round 2. This approach is called the randomization test or permutation test. As with Fisher’s exact test, randomization tests avoid the need to approximate the sampling distribution with a normal distribution. You will be seeing randomization tests more often as computing becomes more powerful.

**Looking for a Better Approach**

D24. Describe how to use the randomization test to determine if an average age as large or larger than that of the 28 workers laid off at Westvaco, 49.86 years, can reasonably be attributed to chance. Each student in your class should perform one trial of your simulation.

D25. A stem-and-leaf plot for the ages of the 50 Westvaco employees is shown in Display 12.18. One condition for a t-test is that your data come from a normal distribution. How appropriate is that assumption here? (In what ways is the shape of the distribution different from normal? How important do you consider these departures from normal?)

```
2 | 2 3
  | 5 9
3 | 0 1 1 2 2 3 4
  | 5 7 8
4 | 2 2
  | 7 8 8 8 9
5 | 0 2 3 3 3 4 4 4
  | 5 5 5 5 6 6 6 7 9 9 9
6 | 0 1 1 3 4 4
  | 6 9
```

**Display 12.18** A stem-and-leaf plot of the ages of the 50 Westvaco employees.

**The End of the Story**

The planning that led to the layoffs at Westvaco took place in several stages. In the first stage, the head of the engineering department made a list of 11 employees to lay off. His boss reviewed the list and decided it was too short: They needed to reduce the size of their workforce even further. The department head added a second group of people to the list and checked again with his boss—still too few. He added a third group of names, then a fourth, and finally one more person in the fifth round of planning. Display 1.1 on page 5 shows this information in the column headed Round: An entry of 1 means “chosen for layoff in Round 1 of the planning,” and similarly for 2, 3, 4, and 5. An entry of 0 means “Retained.”
As it turned out, older employees fared much worse in the earlier rounds of the planning than in the later rounds, as you can see in Display 12.19.

### Display 12.19 Breakdown of layoffs by round and age group.

<table>
<thead>
<tr>
<th>Terminated in Round</th>
<th>Percentage Terminated</th>
<th>Terminated in Round</th>
<th>Percentage Terminated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td>Total</td>
</tr>
<tr>
<td>Under 50</td>
<td>18</td>
<td>4</td>
<td>22</td>
</tr>
<tr>
<td>50 or Older</td>
<td>12</td>
<td>16</td>
<td>28</td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
<td>20</td>
<td>50</td>
</tr>
</tbody>
</table>

Time plots give another view of the same situation. Each plot in Display 12.20 shows the number of employees remaining at Westvaco after each round of layoffs.

### Display 12.20 A time plot of the reduction in work force by round and age group.

**The End of the Story**

D26. Describe the patterns in Display 12.20. What do they suggest to you?

About all you can ever show by using statistical methods in discrimination cases is that the process doesn’t look like random selection. Statistical analysis alerts us to questionable situations, but it cannot reconstruct the intent of the people who did the laying off. Knowing the intent is crucial because it might be perfectly legal: Perhaps employees in obsolete jobs were the ones picked for termination, and it just happens that the obsolete jobs were held by older employees.

The Westvaco case never got as far as a jury. Just before it was about to go to trial, the two sides agreed on a settlement. Details of such settlements are not public information, so, like many problems based on statistics, this case has no “final answer.”
Comparing Termination Rates for Two Age Groups

P29. Consider two companies: Seniors, Inc., with almost all of its 50 employees age 40 or older, and Youth Enterprises, with roughly half of its 50 employees under age 40. Suppose both companies discriminate against older workers in a layoff. If you do a significance test of the difference between two proportions, will you be more likely to detect the discrimination at Seniors, Inc., or at Youth Enterprises? Explain.

P30. Refer to the stem-and-leaf plot of the ages of the 50 Westvaco employees in Display 12.18 on page 821.
   a. Use it to construct a boxplot of the ages.
   b. What characteristics of the distribution can you see from the boxplot that you could not see from the stem-and-leaf plot? From the stem-and-leaf plot that you could not see from the boxplot?

P31. If you use age 40 as the cutoff for defining your age groups, what percentage of the workers are in the older age group? If you use age 50 instead of age 40, what percentage are in the older age group?

P32. Of the 28 workers who were age 50 or older, 19 were laid off. Use this information to construct a table similar to that in D20, part c, but use age 50 as your cutoff age.

P33. Using your table in P32, state appropriate null and alternative hypotheses and carry out a test of the null hypothesis. What do you conclude if you take the results at face value?

P34. Compare your analyses in D21, in which age 40 was the cutoff age, and in P33, in which age 50 was the cutoff age. Which cutoff value, 40 or 50, leads to stronger evidence of discrimination? Which test do you think is more informative about what actually happened—using age 40 as your cutoff age or using age 50? Explain your reasoning.

P35. According to the U.S. Supreme Court, if you do a statistical test of discrimination, you should reject the null hypothesis if your $P$-value from a one-sided test is less than or equal to 0.025 or if your $P$-value from a two-sided test is less than or equal to 0.05.
   a. According to this standard, should the null hypothesis be rejected in either of the two tests—using age 40 as your cutoff (as in D21), or using age 50 (as in P33)?
   b. Which of these statements correctly completes this phrase: If you use a $P$-value of 0.025 to determine “guilt,” then
      I. 2.5% of “not guilty” companies would be declared “guilty” by the test
      II. 2.5% of “guilty” companies would be declared “not guilty” by the test
   c. Is 2.5% the probability of a Type I or a Type II error?
   d. Describe a Type II error in this situation.

Looking for a Better Approach

P36. Construct a back-to-back stemplot of the ages of the Westvaco workers laid off and those retained. Compare the distributions. (Use split stems: 2 | for 20–24, 3 | for 25–29, and so on.)

P37. Use your stemplot from P36 to construct side-by-side boxplots.

P38. Display 12.21 shows the means and standard deviations of the ages for the laid-off and retained workers. Use the summary data to carry out a two-sample $t$-test. (Again, don't worry about conditions for now. Should you use a one-sided test or a two-sided test?)

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Mean Age</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laid off</td>
<td>28</td>
<td>49.86</td>
<td>13.40</td>
</tr>
<tr>
<td>Retained</td>
<td>22</td>
<td>46.18</td>
<td>11.00</td>
</tr>
<tr>
<td>All Employees</td>
<td>50</td>
<td>48.24</td>
<td>12.42</td>
</tr>
</tbody>
</table>

Display 12.21 Means and standard deviations of ages for laid-off and retained employees.

P39. Compare your two-sample $t$-test in P38 with the randomization test in D24. How do the hypotheses, the conditions you need to check, the sampling distributions, and the conclusions differ?
### TABLE A  Standard Normal Probabilities

<table>
<thead>
<tr>
<th>$z$</th>
<th>.00</th>
<th>.01</th>
<th>.02</th>
<th>.03</th>
<th>.04</th>
<th>.05</th>
<th>.06</th>
<th>.07</th>
<th>.08</th>
<th>.09</th>
</tr>
</thead>
<tbody>
<tr>
<td>$-3.8$</td>
<td>.0001</td>
<td>.0001</td>
<td>.0001</td>
<td>.0001</td>
<td>.0001</td>
<td>.0001</td>
<td>.0001</td>
<td>.0001</td>
<td>.0001</td>
<td>.0001</td>
</tr>
<tr>
<td>$-3.7$</td>
<td>.0001</td>
<td>.0001</td>
<td>.0001</td>
<td>.0001</td>
<td>.0001</td>
<td>.0001</td>
<td>.0001</td>
<td>.0001</td>
<td>.0001</td>
<td>.0001</td>
</tr>
<tr>
<td>$-3.6$</td>
<td>.0002</td>
<td>.0002</td>
<td>.0001</td>
<td>.0001</td>
<td>.0001</td>
<td>.0001</td>
<td>.0001</td>
<td>.0001</td>
<td>.0001</td>
<td>.0001</td>
</tr>
<tr>
<td>$-3.5$</td>
<td>.0002</td>
<td>.0002</td>
<td>.0002</td>
<td>.0002</td>
<td>.0002</td>
<td>.0002</td>
<td>.0002</td>
<td>.0002</td>
<td>.0002</td>
<td>.0002</td>
</tr>
<tr>
<td>$-3.3$</td>
<td>.0005</td>
<td>.0005</td>
<td>.0005</td>
<td>.0004</td>
<td>.0004</td>
<td>.0004</td>
<td>.0004</td>
<td>.0004</td>
<td>.0004</td>
<td>.0004</td>
</tr>
<tr>
<td>$-3.2$</td>
<td>.0007</td>
<td>.0007</td>
<td>.0006</td>
<td>.0006</td>
<td>.0006</td>
<td>.0006</td>
<td>.0006</td>
<td>.0006</td>
<td>.0006</td>
<td>.0006</td>
</tr>
<tr>
<td>$-3.1$</td>
<td>.0010</td>
<td>.0009</td>
<td>.0009</td>
<td>.0008</td>
<td>.0008</td>
<td>.0008</td>
<td>.0008</td>
<td>.0008</td>
<td>.0008</td>
<td>.0008</td>
</tr>
<tr>
<td>$-2.9$</td>
<td>.0019</td>
<td>.0018</td>
<td>.0017</td>
<td>.0016</td>
<td>.0016</td>
<td>.0016</td>
<td>.0016</td>
<td>.0016</td>
<td>.0016</td>
<td>.0016</td>
</tr>
<tr>
<td>$-2.8$</td>
<td>.0026</td>
<td>.0025</td>
<td>.0024</td>
<td>.0023</td>
<td>.0023</td>
<td>.0022</td>
<td>.0022</td>
<td>.0022</td>
<td>.0022</td>
<td>.0022</td>
</tr>
<tr>
<td>$-2.7$</td>
<td>.0035</td>
<td>.0034</td>
<td>.0033</td>
<td>.0032</td>
<td>.0031</td>
<td>.0030</td>
<td>.0029</td>
<td>.0029</td>
<td>.0029</td>
<td>.0029</td>
</tr>
<tr>
<td>$-2.6$</td>
<td>.0047</td>
<td>.0045</td>
<td>.0044</td>
<td>.0043</td>
<td>.0041</td>
<td>.0040</td>
<td>.0039</td>
<td>.0039</td>
<td>.0039</td>
<td>.0039</td>
</tr>
<tr>
<td>$-2.5$</td>
<td>.0062</td>
<td>.0060</td>
<td>.0059</td>
<td>.0057</td>
<td>.0055</td>
<td>.0054</td>
<td>.0052</td>
<td>.0052</td>
<td>.0052</td>
<td>.0052</td>
</tr>
<tr>
<td>$-2.4$</td>
<td>.0082</td>
<td>.0080</td>
<td>.0078</td>
<td>.0075</td>
<td>.0073</td>
<td>.0071</td>
<td>.0069</td>
<td>.0068</td>
<td>.0066</td>
<td>.0064</td>
</tr>
<tr>
<td>$-2.3$</td>
<td>.0107</td>
<td>.0104</td>
<td>.0102</td>
<td>.0099</td>
<td>.0096</td>
<td>.0094</td>
<td>.0091</td>
<td>.0089</td>
<td>.0087</td>
<td>.0084</td>
</tr>
<tr>
<td>$-2.2$</td>
<td>.0139</td>
<td>.0136</td>
<td>.0132</td>
<td>.0129</td>
<td>.0125</td>
<td>.0122</td>
<td>.0119</td>
<td>.0116</td>
<td>.0113</td>
<td>.0110</td>
</tr>
<tr>
<td>$-2.1$</td>
<td>.0179</td>
<td>.0174</td>
<td>.0170</td>
<td>.0166</td>
<td>.0162</td>
<td>.0158</td>
<td>.0154</td>
<td>.0150</td>
<td>.0146</td>
<td>.0143</td>
</tr>
<tr>
<td>$-2.0$</td>
<td>.0228</td>
<td>.0222</td>
<td>.0217</td>
<td>.0212</td>
<td>.0207</td>
<td>.0202</td>
<td>.0197</td>
<td>.0192</td>
<td>.0188</td>
<td>.0183</td>
</tr>
<tr>
<td>$-1.9$</td>
<td>.0287</td>
<td>.0281</td>
<td>.0274</td>
<td>.0268</td>
<td>.0262</td>
<td>.0256</td>
<td>.0250</td>
<td>.0244</td>
<td>.0239</td>
<td>.0233</td>
</tr>
<tr>
<td>$-1.8$</td>
<td>.0359</td>
<td>.0351</td>
<td>.0344</td>
<td>.0336</td>
<td>.0329</td>
<td>.0322</td>
<td>.0314</td>
<td>.0307</td>
<td>.0301</td>
<td>.0294</td>
</tr>
<tr>
<td>$-1.7$</td>
<td>.0446</td>
<td>.0436</td>
<td>.0427</td>
<td>.0418</td>
<td>.0409</td>
<td>.0401</td>
<td>.0392</td>
<td>.0384</td>
<td>.0375</td>
<td>.0367</td>
</tr>
<tr>
<td>$-1.6$</td>
<td>.0548</td>
<td>.0537</td>
<td>.0526</td>
<td>.0516</td>
<td>.0505</td>
<td>.0495</td>
<td>.0485</td>
<td>.0475</td>
<td>.0465</td>
<td>.0455</td>
</tr>
<tr>
<td>$-1.5$</td>
<td>.0668</td>
<td>.0655</td>
<td>.0643</td>
<td>.0630</td>
<td>.0618</td>
<td>.0606</td>
<td>.0594</td>
<td>.0582</td>
<td>.0571</td>
<td>.0559</td>
</tr>
<tr>
<td>$-1.4$</td>
<td>.0808</td>
<td>.0793</td>
<td>.0778</td>
<td>.0764</td>
<td>.0749</td>
<td>.0735</td>
<td>.0721</td>
<td>.0708</td>
<td>.0694</td>
<td>.0681</td>
</tr>
<tr>
<td>$-1.3$</td>
<td>.0968</td>
<td>.0951</td>
<td>.0934</td>
<td>.0918</td>
<td>.0901</td>
<td>.0885</td>
<td>.0869</td>
<td>.0853</td>
<td>.0838</td>
<td>.0823</td>
</tr>
<tr>
<td>$-1.2$</td>
<td>.1151</td>
<td>.1131</td>
<td>.1112</td>
<td>.1093</td>
<td>.1075</td>
<td>.1056</td>
<td>.1038</td>
<td>.1020</td>
<td>.1003</td>
<td>.0985</td>
</tr>
<tr>
<td>$-1.1$</td>
<td>.1357</td>
<td>.1335</td>
<td>.1314</td>
<td>.1292</td>
<td>.1271</td>
<td>.1251</td>
<td>.1230</td>
<td>.1210</td>
<td>.1190</td>
<td>.1170</td>
</tr>
<tr>
<td>$-1.0$</td>
<td>.1587</td>
<td>.1562</td>
<td>.1539</td>
<td>.1515</td>
<td>.1492</td>
<td>.1469</td>
<td>.1446</td>
<td>.1423</td>
<td>.1401</td>
<td>.1379</td>
</tr>
<tr>
<td>$-0.9$</td>
<td>.1841</td>
<td>.1814</td>
<td>.1788</td>
<td>.1762</td>
<td>.1736</td>
<td>.1711</td>
<td>.1685</td>
<td>.1660</td>
<td>.1635</td>
<td>.1611</td>
</tr>
<tr>
<td>$-0.7$</td>
<td>.2420</td>
<td>.2389</td>
<td>.2358</td>
<td>.2327</td>
<td>.2296</td>
<td>.2266</td>
<td>.2236</td>
<td>.2206</td>
<td>.2177</td>
<td>.2148</td>
</tr>
<tr>
<td>$-0.6$</td>
<td>.2743</td>
<td>.2709</td>
<td>.2676</td>
<td>.2643</td>
<td>.2611</td>
<td>.2578</td>
<td>.2546</td>
<td>.2514</td>
<td>.2483</td>
<td>.2451</td>
</tr>
<tr>
<td>$-0.5$</td>
<td>.3085</td>
<td>.3050</td>
<td>.3015</td>
<td>.2981</td>
<td>.2946</td>
<td>.2912</td>
<td>.2877</td>
<td>.2843</td>
<td>.2810</td>
<td>.2776</td>
</tr>
<tr>
<td>$-0.4$</td>
<td>.3446</td>
<td>.3409</td>
<td>.3372</td>
<td>.3336</td>
<td>.3300</td>
<td>.3264</td>
<td>.3228</td>
<td>.3192</td>
<td>.3156</td>
<td>.3121</td>
</tr>
<tr>
<td>$-0.3$</td>
<td>.3821</td>
<td>.3783</td>
<td>.3745</td>
<td>.3707</td>
<td>.3669</td>
<td>.3632</td>
<td>.3594</td>
<td>.3557</td>
<td>.3520</td>
<td>.3483</td>
</tr>
<tr>
<td>$-0.2$</td>
<td>.4207</td>
<td>.4168</td>
<td>.4129</td>
<td>.4090</td>
<td>.4052</td>
<td>.4013</td>
<td>.3974</td>
<td>.3936</td>
<td>.3897</td>
<td>.3859</td>
</tr>
<tr>
<td>$-0.1$</td>
<td>.4602</td>
<td>.4562</td>
<td>.4522</td>
<td>.4483</td>
<td>.4443</td>
<td>.4404</td>
<td>.4364</td>
<td>.4325</td>
<td>.4286</td>
<td>.4247</td>
</tr>
<tr>
<td>$0.0$</td>
<td>.5000</td>
<td>.4960</td>
<td>.4920</td>
<td>.4880</td>
<td>.4840</td>
<td>.4801</td>
<td>.4761</td>
<td>.4721</td>
<td>.4681</td>
<td>.4641</td>
</tr>
</tbody>
</table>
Table entry for $z$ is the probability lying below $z$.

### TABLE A  Standard Normal Probabilities (continued)

<table>
<thead>
<tr>
<th>$z$</th>
<th>.00</th>
<th>.01</th>
<th>.02</th>
<th>.03</th>
<th>.04</th>
<th>.05</th>
<th>.06</th>
<th>.07</th>
<th>.08</th>
<th>.09</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>.5000</td>
<td>.5040</td>
<td>.5080</td>
<td>.5120</td>
<td>.5160</td>
<td>.5199</td>
<td>.5239</td>
<td>.5279</td>
<td>.5319</td>
<td>.5359</td>
</tr>
<tr>
<td>0.1</td>
<td>.5398</td>
<td>.5438</td>
<td>.5478</td>
<td>.5517</td>
<td>.5557</td>
<td>.5596</td>
<td>.5636</td>
<td>.5675</td>
<td>.5714</td>
<td>.5753</td>
</tr>
<tr>
<td>0.2</td>
<td>.5793</td>
<td>.5832</td>
<td>.5871</td>
<td>.5910</td>
<td>.5948</td>
<td>.5987</td>
<td>.6026</td>
<td>.6064</td>
<td>.6103</td>
<td>.6141</td>
</tr>
<tr>
<td>0.3</td>
<td>.6179</td>
<td>.6217</td>
<td>.6255</td>
<td>.6293</td>
<td>.6331</td>
<td>.6368</td>
<td>.6406</td>
<td>.6443</td>
<td>.6480</td>
<td>.6517</td>
</tr>
<tr>
<td>0.4</td>
<td>.6554</td>
<td>.6591</td>
<td>.6628</td>
<td>.6664</td>
<td>.6700</td>
<td>.6736</td>
<td>.6772</td>
<td>.6808</td>
<td>.6844</td>
<td>.6879</td>
</tr>
<tr>
<td>0.5</td>
<td>.6915</td>
<td>.6950</td>
<td>.6985</td>
<td>.7020</td>
<td>.7054</td>
<td>.7088</td>
<td>.7123</td>
<td>.7157</td>
<td>.7190</td>
<td>.7224</td>
</tr>
<tr>
<td>0.6</td>
<td>.7257</td>
<td>.7291</td>
<td>.7324</td>
<td>.7357</td>
<td>.7389</td>
<td>.7422</td>
<td>.7454</td>
<td>.7486</td>
<td>.7517</td>
<td>.7549</td>
</tr>
<tr>
<td>0.7</td>
<td>.7580</td>
<td>.7611</td>
<td>.7642</td>
<td>.7673</td>
<td>.7704</td>
<td>.7734</td>
<td>.7764</td>
<td>.7794</td>
<td>.7823</td>
<td>.7852</td>
</tr>
<tr>
<td>0.8</td>
<td>.7881</td>
<td>.7910</td>
<td>.7939</td>
<td>.7967</td>
<td>.7995</td>
<td>.8023</td>
<td>.8051</td>
<td>.8078</td>
<td>.8106</td>
<td>.8133</td>
</tr>
<tr>
<td>0.9</td>
<td>.8159</td>
<td>.8186</td>
<td>.8212</td>
<td>.8238</td>
<td>.8264</td>
<td>.8289</td>
<td>.8315</td>
<td>.8340</td>
<td>.8365</td>
<td>.8389</td>
</tr>
<tr>
<td>1.0</td>
<td>.8413</td>
<td>.8438</td>
<td>.8461</td>
<td>.8485</td>
<td>.8508</td>
<td>.8531</td>
<td>.8554</td>
<td>.8577</td>
<td>.8599</td>
<td>.8621</td>
</tr>
<tr>
<td>1.1</td>
<td>.8643</td>
<td>.8665</td>
<td>.8686</td>
<td>.8708</td>
<td>.8729</td>
<td>.8749</td>
<td>.8770</td>
<td>.8790</td>
<td>.8810</td>
<td>.8830</td>
</tr>
<tr>
<td>1.2</td>
<td>.8849</td>
<td>.8869</td>
<td>.8888</td>
<td>.8907</td>
<td>.8925</td>
<td>.8944</td>
<td>.8962</td>
<td>.8980</td>
<td>.8997</td>
<td>.9015</td>
</tr>
<tr>
<td>1.3</td>
<td>.9032</td>
<td>.9049</td>
<td>.9066</td>
<td>.9082</td>
<td>.9099</td>
<td>.9115</td>
<td>.9131</td>
<td>.9147</td>
<td>.9162</td>
<td>.9177</td>
</tr>
<tr>
<td>1.4</td>
<td>.9192</td>
<td>.9207</td>
<td>.9222</td>
<td>.9236</td>
<td>.9251</td>
<td>.9265</td>
<td>.9279</td>
<td>.9292</td>
<td>.9306</td>
<td>.9319</td>
</tr>
<tr>
<td>1.5</td>
<td>.9332</td>
<td>.9345</td>
<td>.9357</td>
<td>.9370</td>
<td>.9382</td>
<td>.9394</td>
<td>.9406</td>
<td>.9418</td>
<td>.9429</td>
<td>.9441</td>
</tr>
<tr>
<td>1.6</td>
<td>.9452</td>
<td>.9463</td>
<td>.9474</td>
<td>.9484</td>
<td>.9495</td>
<td>.9505</td>
<td>.9515</td>
<td>.9525</td>
<td>.9535</td>
<td>.9545</td>
</tr>
<tr>
<td>1.7</td>
<td>.9554</td>
<td>.9564</td>
<td>.9573</td>
<td>.9582</td>
<td>.9591</td>
<td>.9599</td>
<td>.9608</td>
<td>.9616</td>
<td>.9625</td>
<td>.9633</td>
</tr>
<tr>
<td>1.8</td>
<td>.9641</td>
<td>.9649</td>
<td>.9656</td>
<td>.9664</td>
<td>.9671</td>
<td>.9678</td>
<td>.9686</td>
<td>.9693</td>
<td>.9699</td>
<td>.9706</td>
</tr>
<tr>
<td>1.9</td>
<td>.9713</td>
<td>.9719</td>
<td>.9726</td>
<td>.9732</td>
<td>.9738</td>
<td>.9744</td>
<td>.9750</td>
<td>.9756</td>
<td>.9761</td>
<td>.9767</td>
</tr>
<tr>
<td>2.0</td>
<td>.9772</td>
<td>.9778</td>
<td>.9783</td>
<td>.9788</td>
<td>.9793</td>
<td>.9798</td>
<td>.9803</td>
<td>.9808</td>
<td>.9812</td>
<td>.9817</td>
</tr>
<tr>
<td>2.1</td>
<td>.9821</td>
<td>.9826</td>
<td>.9830</td>
<td>.9834</td>
<td>.9838</td>
<td>.9842</td>
<td>.9846</td>
<td>.9850</td>
<td>.9854</td>
<td>.9857</td>
</tr>
<tr>
<td>2.2</td>
<td>.9861</td>
<td>.9864</td>
<td>.9868</td>
<td>.9871</td>
<td>.9875</td>
<td>.9878</td>
<td>.9881</td>
<td>.9884</td>
<td>.9887</td>
<td>.9890</td>
</tr>
<tr>
<td>2.3</td>
<td>.9893</td>
<td>.9896</td>
<td>.9898</td>
<td>.9901</td>
<td>.9904</td>
<td>.9906</td>
<td>.9909</td>
<td>.9911</td>
<td>.9913</td>
<td>.9916</td>
</tr>
<tr>
<td>2.4</td>
<td>.9918</td>
<td>.9920</td>
<td>.9922</td>
<td>.9925</td>
<td>.9927</td>
<td>.9929</td>
<td>.9931</td>
<td>.9932</td>
<td>.9934</td>
<td>.9936</td>
</tr>
<tr>
<td>2.5</td>
<td>.9938</td>
<td>.9940</td>
<td>.9941</td>
<td>.9943</td>
<td>.9945</td>
<td>.9946</td>
<td>.9948</td>
<td>.9949</td>
<td>.9951</td>
<td>.9952</td>
</tr>
<tr>
<td>2.6</td>
<td>.9953</td>
<td>.9955</td>
<td>.9956</td>
<td>.9957</td>
<td>.9959</td>
<td>.9960</td>
<td>.9961</td>
<td>.9962</td>
<td>.9963</td>
<td>.9964</td>
</tr>
<tr>
<td>2.7</td>
<td>.9965</td>
<td>.9966</td>
<td>.9967</td>
<td>.9968</td>
<td>.9969</td>
<td>.9970</td>
<td>.9971</td>
<td>.9972</td>
<td>.9973</td>
<td>.9974</td>
</tr>
<tr>
<td>2.8</td>
<td>.9974</td>
<td>.9975</td>
<td>.9976</td>
<td>.9977</td>
<td>.9977</td>
<td>.9978</td>
<td>.9979</td>
<td>.9979</td>
<td>.9980</td>
<td>.9981</td>
</tr>
<tr>
<td>2.9</td>
<td>.9981</td>
<td>.9982</td>
<td>.9982</td>
<td>.9983</td>
<td>.9984</td>
<td>.9984</td>
<td>.9985</td>
<td>.9985</td>
<td>.9986</td>
<td>.9986</td>
</tr>
<tr>
<td>3.0</td>
<td>.9987</td>
<td>.9987</td>
<td>.9987</td>
<td>.9988</td>
<td>.9988</td>
<td>.9989</td>
<td>.9989</td>
<td>.9989</td>
<td>.9989</td>
<td>.9989</td>
</tr>
<tr>
<td>3.1</td>
<td>.9990</td>
<td>.9991</td>
<td>.9991</td>
<td>.9991</td>
<td>.9992</td>
<td>.9992</td>
<td>.9992</td>
<td>.9992</td>
<td>.9993</td>
<td>.9993</td>
</tr>
<tr>
<td>3.2</td>
<td>.9993</td>
<td>.9993</td>
<td>.9994</td>
<td>.9994</td>
<td>.9994</td>
<td>.9994</td>
<td>.9994</td>
<td>.9994</td>
<td>.9995</td>
<td>.9995</td>
</tr>
<tr>
<td>3.3</td>
<td>.9995</td>
<td>.9995</td>
<td>.9995</td>
<td>.9996</td>
<td>.9996</td>
<td>.9996</td>
<td>.9996</td>
<td>.9996</td>
<td>.9996</td>
<td>.9997</td>
</tr>
<tr>
<td>3.4</td>
<td>.9997</td>
<td>.9997</td>
<td>.9997</td>
<td>.9997</td>
<td>.9997</td>
<td>.9997</td>
<td>.9997</td>
<td>.9997</td>
<td>.9997</td>
<td>.9998</td>
</tr>
<tr>
<td>3.5</td>
<td>.9998</td>
<td>.9998</td>
<td>.9998</td>
<td>.9998</td>
<td>.9998</td>
<td>.9998</td>
<td>.9998</td>
<td>.9998</td>
<td>.9998</td>
<td>.9998</td>
</tr>
<tr>
<td>3.6</td>
<td>.9998</td>
<td>.9998</td>
<td>.9999</td>
<td>.9999</td>
<td>.9999</td>
<td>.9999</td>
<td>.9999</td>
<td>.9999</td>
<td>.9999</td>
<td>.9999</td>
</tr>
<tr>
<td>3.7</td>
<td>.9999</td>
<td>.9999</td>
<td>.9999</td>
<td>.9999</td>
<td>.9999</td>
<td>.9999</td>
<td>.9999</td>
<td>.9999</td>
<td>.9999</td>
<td>.9999</td>
</tr>
<tr>
<td>3.8</td>
<td>.9999</td>
<td>.9999</td>
<td>.9999</td>
<td>.9999</td>
<td>.9999</td>
<td>.9999</td>
<td>.9999</td>
<td>.9999</td>
<td>.9999</td>
<td>.9999</td>
</tr>
</tbody>
</table>
Table entry for \( p \) and \( C \) is the value of \( t^* \) with probability \( p \) lying above it and probability \( C \) lying between \(-t^*\) and \( t^*\).

### TABLE B  \( t \)-Distribution Critical Values

<table>
<thead>
<tr>
<th>df</th>
<th>.25</th>
<th>.20</th>
<th>.15</th>
<th>.10</th>
<th>.05</th>
<th>.025</th>
<th>.02</th>
<th>.01</th>
<th>.005</th>
<th>.0025</th>
<th>.001</th>
<th>.0005</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.000</td>
<td>1.376</td>
<td>1.963</td>
<td>3.078</td>
<td>6.314</td>
<td>12.71</td>
<td>15.89</td>
<td>31.82</td>
<td>63.66</td>
<td>127.3</td>
<td>318.3</td>
<td>636.6</td>
</tr>
<tr>
<td>10</td>
<td>.676</td>
<td>.825</td>
<td>1.030</td>
<td>1.189</td>
<td>1.733</td>
<td>2.120</td>
<td>2.258</td>
<td>2.718</td>
<td>2.920</td>
<td>3.499</td>
<td>4.290</td>
<td>6.302</td>
</tr>
<tr>
<td>20</td>
<td>.667</td>
<td>.797</td>
<td>1.004</td>
<td>1.118</td>
<td>1.636</td>
<td>2.000</td>
<td>2.101</td>
<td>2.565</td>
<td>2.704</td>
<td>3.144</td>
<td>3.833</td>
<td>5.448</td>
</tr>
<tr>
<td>50</td>
<td>.648</td>
<td>.764</td>
<td>.965</td>
<td>1.056</td>
<td>1.533</td>
<td>1.915</td>
<td>2.012</td>
<td>2.423</td>
<td>2.528</td>
<td>2.878</td>
<td>3.499</td>
<td>5.043</td>
</tr>
<tr>
<td>100</td>
<td>.636</td>
<td>.748</td>
<td>.945</td>
<td>1.033</td>
<td>1.486</td>
<td>1.860</td>
<td>1.984</td>
<td>2.364</td>
<td>2.433</td>
<td>2.681</td>
<td>3.250</td>
<td>4.800</td>
</tr>
<tr>
<td>500</td>
<td>.625</td>
<td>.732</td>
<td>.931</td>
<td>1.021</td>
<td>1.440</td>
<td>1.816</td>
<td>1.940</td>
<td>2.310</td>
<td>2.370</td>
<td>2.551</td>
<td>3.078</td>
<td>4.541</td>
</tr>
<tr>
<td>1000</td>
<td>.620</td>
<td>.726</td>
<td>.925</td>
<td>1.016</td>
<td>1.420</td>
<td>1.782</td>
<td>1.900</td>
<td>2.262</td>
<td>2.321</td>
<td>2.492</td>
<td>2.947</td>
<td>4.319</td>
</tr>
</tbody>
</table>

Confidence level \( C \)
Table entry for $p$ is the value of $\chi^2$ with probability $p$ lying above it.

**TABLE C  $\chi^2$ Critical Values**

<table>
<thead>
<tr>
<th>df</th>
<th>.25</th>
<th>.20</th>
<th>.15</th>
<th>.10</th>
<th>.05</th>
<th>.025</th>
<th>.02</th>
<th>.01</th>
<th>.005</th>
<th>.0025</th>
<th>.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.32</td>
<td>1.64</td>
<td>2.07</td>
<td>2.71</td>
<td>3.84</td>
<td>5.02</td>
<td>6.63</td>
<td>7.88</td>
<td>9.14</td>
<td>10.83</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2.77</td>
<td>3.22</td>
<td>3.79</td>
<td>4.61</td>
<td>5.99</td>
<td>7.38</td>
<td>8.62</td>
<td>9.22</td>
<td>9.78</td>
<td>10.60</td>
<td>11.88</td>
</tr>
<tr>
<td>3</td>
<td>4.11</td>
<td>4.64</td>
<td>5.32</td>
<td>6.25</td>
<td>7.81</td>
<td>9.35</td>
<td>11.07</td>
<td>12.83</td>
<td>13.33</td>
<td>15.09</td>
<td>16.81</td>
</tr>
<tr>
<td>6</td>
<td>7.84</td>
<td>8.56</td>
<td>9.45</td>
<td>10.64</td>
<td>12.53</td>
<td>14.45</td>
<td>15.03</td>
<td>16.81</td>
<td>18.55</td>
<td>20.25</td>
<td>22.46</td>
</tr>
<tr>
<td>8</td>
<td>10.22</td>
<td>11.03</td>
<td>12.03</td>
<td>13.36</td>
<td>15.51</td>
<td>17.53</td>
<td>19.17</td>
<td>21.09</td>
<td>22.95</td>
<td>24.63</td>
<td>26.92</td>
</tr>
<tr>
<td>11</td>
<td>13.70</td>
<td>14.63</td>
<td>15.77</td>
<td>17.29</td>
<td>19.68</td>
<td>21.92</td>
<td>23.62</td>
<td>25.76</td>
<td>27.87</td>
<td>30.32</td>
<td>32.91</td>
</tr>
<tr>
<td>12</td>
<td>14.85</td>
<td>15.81</td>
<td>16.99</td>
<td>18.55</td>
<td>21.03</td>
<td>23.34</td>
<td>25.05</td>
<td>27.28</td>
<td>29.54</td>
<td>32.12</td>
<td>34.75</td>
</tr>
<tr>
<td>13</td>
<td>15.98</td>
<td>16.98</td>
<td>18.20</td>
<td>20.67</td>
<td>23.19</td>
<td>25.57</td>
<td>27.86</td>
<td>30.19</td>
<td>32.67</td>
<td>35.28</td>
<td>37.93</td>
</tr>
<tr>
<td>14</td>
<td>17.12</td>
<td>18.15</td>
<td>19.41</td>
<td>21.97</td>
<td>24.36</td>
<td>26.87</td>
<td>29.72</td>
<td>32.14</td>
<td>34.83</td>
<td>37.93</td>
<td>40.70</td>
</tr>
<tr>
<td>15</td>
<td>18.25</td>
<td>19.31</td>
<td>20.60</td>
<td>23.12</td>
<td>25.70</td>
<td>28.17</td>
<td>30.88</td>
<td>33.33</td>
<td>36.21</td>
<td>39.32</td>
<td>42.31</td>
</tr>
<tr>
<td>16</td>
<td>19.37</td>
<td>20.47</td>
<td>21.79</td>
<td>24.36</td>
<td>26.60</td>
<td>28.95</td>
<td>31.42</td>
<td>34.00</td>
<td>37.16</td>
<td>40.66</td>
<td>43.95</td>
</tr>
<tr>
<td>17</td>
<td>20.49</td>
<td>21.61</td>
<td>22.98</td>
<td>25.79</td>
<td>28.43</td>
<td>30.89</td>
<td>33.24</td>
<td>35.80</td>
<td>39.12</td>
<td>42.66</td>
<td>46.36</td>
</tr>
<tr>
<td>18</td>
<td>21.60</td>
<td>22.76</td>
<td>24.16</td>
<td>25.99</td>
<td>28.87</td>
<td>31.53</td>
<td>34.61</td>
<td>37.87</td>
<td>40.79</td>
<td>44.20</td>
<td>47.39</td>
</tr>
<tr>
<td>19</td>
<td>22.72</td>
<td>23.90</td>
<td>25.33</td>
<td>27.20</td>
<td>30.14</td>
<td>32.85</td>
<td>35.69</td>
<td>38.85</td>
<td>41.62</td>
<td>45.00</td>
<td>48.08</td>
</tr>
<tr>
<td>20</td>
<td>23.83</td>
<td>25.04</td>
<td>26.50</td>
<td>28.41</td>
<td>31.41</td>
<td>34.17</td>
<td>37.57</td>
<td>40.00</td>
<td>42.85</td>
<td>46.49</td>
<td>50.02</td>
</tr>
<tr>
<td>21</td>
<td>24.93</td>
<td>26.17</td>
<td>27.66</td>
<td>29.62</td>
<td>32.67</td>
<td>35.48</td>
<td>38.93</td>
<td>41.40</td>
<td>43.78</td>
<td>47.92</td>
<td>51.97</td>
</tr>
<tr>
<td>22</td>
<td>26.04</td>
<td>27.30</td>
<td>28.82</td>
<td>30.81</td>
<td>33.92</td>
<td>36.78</td>
<td>37.66</td>
<td>40.29</td>
<td>42.80</td>
<td>46.38</td>
<td>51.66</td>
</tr>
<tr>
<td>23</td>
<td>27.14</td>
<td>28.43</td>
<td>29.98</td>
<td>32.01</td>
<td>35.17</td>
<td>38.08</td>
<td>38.97</td>
<td>41.64</td>
<td>44.18</td>
<td>47.73</td>
<td>52.27</td>
</tr>
<tr>
<td>24</td>
<td>28.24</td>
<td>29.55</td>
<td>31.13</td>
<td>33.20</td>
<td>36.42</td>
<td>39.36</td>
<td>40.27</td>
<td>42.98</td>
<td>45.56</td>
<td>48.98</td>
<td>53.68</td>
</tr>
<tr>
<td>25</td>
<td>29.34</td>
<td>30.68</td>
<td>32.28</td>
<td>34.38</td>
<td>37.65</td>
<td>40.65</td>
<td>41.57</td>
<td>44.31</td>
<td>47.93</td>
<td>51.49</td>
<td>55.08</td>
</tr>
<tr>
<td>26</td>
<td>30.43</td>
<td>31.79</td>
<td>33.43</td>
<td>35.56</td>
<td>38.89</td>
<td>41.92</td>
<td>42.86</td>
<td>45.64</td>
<td>48.29</td>
<td>52.03</td>
<td>55.40</td>
</tr>
<tr>
<td>27</td>
<td>31.53</td>
<td>32.91</td>
<td>34.57</td>
<td>36.74</td>
<td>40.11</td>
<td>43.19</td>
<td>44.14</td>
<td>46.96</td>
<td>49.64</td>
<td>53.22</td>
<td>57.48</td>
</tr>
<tr>
<td>28</td>
<td>32.62</td>
<td>34.03</td>
<td>35.71</td>
<td>37.92</td>
<td>41.34</td>
<td>44.46</td>
<td>45.42</td>
<td>48.28</td>
<td>50.99</td>
<td>54.59</td>
<td>59.89</td>
</tr>
<tr>
<td>29</td>
<td>33.71</td>
<td>35.14</td>
<td>36.85</td>
<td>39.09</td>
<td>42.56</td>
<td>45.72</td>
<td>47.69</td>
<td>49.59</td>
<td>52.34</td>
<td>55.97</td>
<td>63.30</td>
</tr>
<tr>
<td>30</td>
<td>34.80</td>
<td>36.25</td>
<td>37.99</td>
<td>40.26</td>
<td>43.77</td>
<td>46.98</td>
<td>47.96</td>
<td>50.89</td>
<td>53.67</td>
<td>56.33</td>
<td>59.70</td>
</tr>
<tr>
<td>40</td>
<td>45.62</td>
<td>47.27</td>
<td>49.24</td>
<td>51.81</td>
<td>55.76</td>
<td>59.34</td>
<td>60.44</td>
<td>63.69</td>
<td>66.77</td>
<td>69.70</td>
<td>73.40</td>
</tr>
<tr>
<td>50</td>
<td>56.33</td>
<td>53.16</td>
<td>60.35</td>
<td>63.17</td>
<td>67.50</td>
<td>71.42</td>
<td>72.61</td>
<td>76.15</td>
<td>79.49</td>
<td>82.66</td>
<td>86.66</td>
</tr>
<tr>
<td>60</td>
<td>66.98</td>
<td>68.97</td>
<td>71.34</td>
<td>74.40</td>
<td>79.08</td>
<td>83.30</td>
<td>84.58</td>
<td>88.38</td>
<td>91.95</td>
<td>95.34</td>
<td>99.61</td>
</tr>
<tr>
<td>80</td>
<td>88.13</td>
<td>90.41</td>
<td>93.11</td>
<td>96.58</td>
<td>101.9</td>
<td>106.6</td>
<td>108.1</td>
<td>112.3</td>
<td>116.3</td>
<td>120.1</td>
<td>124.8</td>
</tr>
<tr>
<td>100</td>
<td>109.1</td>
<td>111.7</td>
<td>114.7</td>
<td>118.5</td>
<td>124.3</td>
<td>129.6</td>
<td>131.1</td>
<td>135.8</td>
<td>140.2</td>
<td>144.3</td>
<td>149.4</td>
</tr>
</tbody>
</table>
TABLE D

Random Digits

Row
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50

828

10097
37542
08422
99019
12807
66065
31060
85269
63573
73796
98520
11805
83452
88685
99594
65481
80124
74350
69916
09893
91499
80336
44104
12550
63606
61196
15474
94557
42481
23523
04493
00549
35963
59808
46058
32179
69234
19565
45155
94864
98086
33185
80951
79752
18633
74029
54178
11664
48324
69074

Appendix: Statistical Tables

32533
04805
68953
02529
99970
74717
10805
77602
32135
45753
17767
05431
99634
40200
67348
17674
35635
99817
26803
20505
14523
94598
81949
73742
49329
90446
45266
28573
16213
78317
52494
97654
15307
08391
85236
00597
61406
41430
14938
31994
24826
16232
00406
49140
32537
43902
45611
49883
77928
94138

76520
64894
19645
09376
80157
34072
45571
02051
05325
03529
14905
39808
06288
86507
87517
17468
17727
77402
66252
14225
68479
26940
85157
11100
16505
26457
95270
67897
97344
73208
75246
64051
26898
45427
01390
87379
20117
01758
19476
36168
45240
41941
96382
71961
98145
77557
80993
52079
31249
87637

13586
74296
09303
70715
36147
76850
82406
65692
47048
64778
68607
27732
98083
58401
64969
50950
08015
77214
29148
68514
27686
36858
47954
02040
34484
47774
79953
54387
08721
89837
33824
88159
09354
26842
92286
25241
45204
75379
07246
10851
28404
50949
70774
28296
06571
32270
37143
84827
64710
91976

34673
24805
23209
38311
64032
36697
35303
68665
90553
35808
22109
50725
13746
36766
91826
58047
45318
43236
36936
46427
46162
70297
32979
12860
40219
51924
59367
54622
16868
68935
45862
96119
33351
83609
77281
05567
15956
40419
43667
34888
44999
89435
20151
69861
31010
97790
05335
59381
02295
35584

54876
24037
02560
31165
36653
36170
42614
74818
57548
34282
40558
68248
70078
67951
08928
76974
22374
00210
87203
56788
83554
34135
26575
74697
52563
33729
83848
44431
48767
91416
51025
63896
35462
49700
44077
07007
60000
21585
94543
81553
08896
48581
23387
02591
24674
17119
12969
71539
36870
04401

80959
20636
15953
88676
98951
65813
86799
73053
28468
60935
60970
29405
18475
90364
93785
73039
21115
45521
76621
96297
94750
53140
57600
96644
43651
65394
82396
91190
03071
26252
61962
54692
77974
13021
93910
86743
18743
66674
59047
01540
39094
88695
25016
74852
05455
52527
56127
09973
32307
10518

09117
10402
34764
74397
16877
39885
07439
85247
28709
20344
93433
24201
40610
76493
61368
57186
78253
64237
13990
78822
89923
33340
40881
89439
77082
59593
10118
42592
12059
29663
79335
82391
50024
24892
83647
17157
92423
36806
90033
35456
73407
41994
25298
20539
61427
58021
19255
33440
57546
21615

39292
00822
35080
04436
12171
11199
23403
18623
83491
35273
50500
52775
68711
29609
23478
40218
14385
96286
94400
54382
37089
42050
22222
28707
07207
42582
33211
92927
25701
05522
65337
23287
90103
78565
70617
85394
97118
84962
20826
05014
35441
37548
94624
00387
77938
80814
36040
88461
15020
01848

74945
91665
33606
27659
76833
29170
09732
88579
25624
88435
73998
67851
77817
11062
34113
16544
53763
02655
56418
14598
20048
82341
06413
25815
31790
60527
59466
45973
46670
82562
12472
29529
39333
20106
42941
11838
96338
85207
69541
51176
31880
73043
61171
59579
91936
51748
90324
23356
09994
76938


### Glossary

#### Symbols

- $b_0$: $y$-intercept of the sample regression line
- $b_1$: slope of the sample regression line
- $E$: margin of error
- $E(X)$: expected value (mean) of the random variable $X$
- $H_0$: null hypothesis
- $H_a$: alternative hypothesis (sometimes $H_1$)
- $IQR$: interquartile range
- $n$: sample size or the number of trials
- $N$: population size
- $p$: population proportion or the probability of a success on any one trial
- $p_0$: hypothesized value of the population proportion
- $\hat{p}$: sample proportion
- $\bar{p}$: estimate of the common value of a population proportion found by combining two samples
- $P(A)$: probability that event $A$ happens
- $P(A \text{ or } B)$: probability that event $A$ happens or event $B$ happens or both (also $P(A \cup B)$)
- $P(A \text{ and } B)$: probability that event $A$ and event $B$ both happen (also $P(A \cap B)$)
- $P(B|A)$: conditional probability that event $B$ happens given that event $A$ happens
- $P(x)$: probability that the random variable $X$ takes on the value $x$ (also $P(X = x)$)
- $Q_1$: first or lower quartile
- $Q_3$: third or upper quartile
- $r$: sample correlation
- $s$: sample standard deviation; in regression, the estimate from the sample of the common variability of $y$ at each value of $x$
- $s^2$: sample variance
- $s_{b_1}$: estimated standard error of the slope of the regression line
- $SD$: standard deviation
- $SE$: standard error
- $SRS$: simple random sample
- $SSE$: sum of squared errors
- $t$: a test statistic using a $t$-distribution
- $\bar{X}$: sample mean
- $x$: observed value of a variable
- $X$: a random variable
- $y$: observed value of a variable
- $\hat{y}$: predicted value of a variable
- $z$: standardized value ($z$-score)
- $z^*$: critical value

#### Greek Letters

- $\alpha$: significance level; probability of a Type I error
- $\beta$: probability of a Type II error
- $\beta_0$: $y$-intercept of the population regression line
- $\beta_1$: slope of the population regression line
- $\varepsilon$: difference of the value of $y$ for a point and the value predicted by the population regression line
- $\mu$: population mean
- $\mu_{\bar{X}}$: mean of the sampling distribution of the sample mean
- $\mu_{y|x}$: mean of the conditional distribution of $y$ given $x$
- $\sigma$: population standard deviation; in regression, the common variability of $y$ at each value of $x$
- $\sigma_{\bar{X}}$: standard error of the mean
**Glossary**

**box-and-whiskers plot**  See boxplot.

**boxplot** (or box-and-whiskers plot) A graphical display of the five-number summary. The “box” extends from the lower quartile to the upper quartile, with a line across it at the median. The “whiskers” run from the quartiles to the minimum and maximum.

**capture rate**  The proportion of confidence intervals produced by a particular method that capture the population parameter. See also confidence level.

**case**  The subject (or unit) on which a measurement is made.

**categorical variable**  A variable that can be grouped into categories, such as “yes” and “no.” Categories sometimes can be ordered, such as “small,” “medium,” and “large.”

**census**  A collection of measurements on all units in the population of interest.

**Central Limit Theorem**  The shape of the sampling distribution of the sample mean becomes more normal as $n$ increases.

**chance model**  See probability model.

**chi-square test of homogeneity**  A chi-square test used to determine whether it is reasonable to believe that when several different populations are broken down into the same categories, they have the same proportion of units in each category.

**chi-square test of independence**  A chi-square test of the hypothesis that two categorical variables measured on the same units are independent of each other in the population.

**clinical trial**  A randomized experiment comparing the effects of medical treatments on human subjects.

**cluster(s)**  On a plot, a group of data “clustering” close to the same value, away from other groups. In sampling, non-overlapping and exhaustive groupings of the units in a population.

**cluster sampling**  Selecting a simple random sample of clusters of units (such as classrooms of students) rather than individual units (students).

**coefficient of determination**  The square of the correlation $r$. Tells the proportion of the total variation in $y$ that can be explained by the relationship with $x$.

**column chart**  A three-dimensional plot of frequencies taken from a two-way table, which depicts those frequencies as heights of columns.
column marginal frequency  The total of all frequencies across row categories for a particular column of a two-way table of frequencies for categorical variables.

comparison group  In an experiment, a group that receives one of the treatments, often the standard treatment.

complement  In probability, the outcomes in the sample space that lie outside an event of interest.

completely randomized design  An experimental design in which treatments are randomly assigned to units without restriction.

conditional distribution of $y$ given $x$  With bivariate data, the distribution of the values of $y$ for a fixed value of $x$.

conditional probability  The notion that a probability can change if you are given additional information. The conditional probability that event $A$ happens given that event $B$ happens is given by $P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$ as long as $P(B) > 0$.

conditional relative frequency  The joint frequency in a column divided by the marginal frequency for that column, or the joint frequency in a row divided by the marginal frequency for that row.

collocation interval  A set of plausible values for a population parameter, any one of which could be used to define a population for which the observed sample statistic would be a reasonably likely outcome.

collocation level  The probability that the method used will give a confidence interval that captures the parameter.

confounding variables (or confounding)  Two variables in an observational study whose effects on the response are impossible to separate.

continuous variable  A quantitative variable that can take on any value in an interval of real numbers.

control group  In an experiment, a group that provides a standard for comparison to evaluate the effectiveness of a treatment; often given a placebo.

convenience sample  A sample in which the units chosen from the population are the units that are easy (convenient) to include, rather than being selected randomly.

correlation  A numerical value between $-1$ and $1$, inclusive, that measures the strength and direction of a linear relationship between two variables.

critical value  The value to which a test statistic is compared in order to decide whether to reject the null hypothesis. Or, the multiplier used in computing the margin of error for a confidence interval.

cumulative percentage plot  See cumulative relative frequency plot.

cumulative relative frequency plot (or cumulative percentage plot)  A plot of ordered pairs in which each value $x$ in the distribution and its cumulative relative frequency, that is, the proportion of all values less than or equal to $x$, are plotted.

data  A set of numbers or observations with a context and drawn from a real-life sample or population.

data analysis  See statistics.

degrees of freedom  The number of freely varying pieces of information on which an estimator is based. For example, when using a sample to estimate the variability in the population, the number of independent deviations from the estimate of center.

dependent events  Events that are not independent.

development  The difference from the mean, $x - \bar{x}$, or from some other measure of center.

disjoint events (or mutually exclusive events)  Events that cannot occur on the same opportunity. If event $A$ and event $B$ are disjoint, $P(A \text{ and } B) = 0$.

distribution, data  The set of values that a variable takes on in a sample or population, together with how frequently each value occurs.

distribution, probability  The set of values that a random variable takes on, together with a means of determining the probability of each value (or interval of values in the case of a continuous distribution).

dot plot  A graphical display that shows the values of a variable along a number line.

double-blind  Describes an experiment in which neither the subjects nor the researcher making the measurements knows which treatment the subjects received.

ecological fallacy  The mistake of using ecological (group-level) correlations to support conclusions about individuals.

event  Any subset of a sample space.

expected value, $\mu_X$ or $E(X)$  The mean of the probability distribution for the random variable $X$.

experimental units  In an experiment, the subjects or objects to which treatments are assigned.

Glossary  831
explanatory variable (or predictor) A variable used to predict (or explain) the value of the response variable. Placed on the x-axis in a regression analysis.

exploratory analysis (or data exploration) An investigation to find patterns in data, using tools such as tables, statistical graphics, and summary statistics to display and summarize distributions.

exponential relationship A relationship between two variables in which the response variable, \( y \), is multiplied by a constant for each unit of increase in the explanatory variable, \( x \). Mathematically, \( y = ab^x \) where \( a \) and \( b \) are constants.

extrapolation Making a prediction when the value of the explanatory variable, \( x \), falls outside the range of the observed data.

factor An explanatory variable, usually categorical, in a randomized experiment or an observational study.

first quartile, \( Q_1 \) See lower quartile.

fitted value See predicted value.

five-number summary A data summary that lists the minimum and maximum values, the median, and the lower and upper quartiles for a data set.

fixed-level test A test in which the null hypothesis is rejected or not rejected based on comparison of the test statistic with the critical value for some predetermined level of significance.

frame See sampling frame.

frequency (or count) The number of times a value occurs in a distribution. With categorical data, the number of units that fall into a specific category.

frequency table A table that gives data values and their frequencies.

Fundamental Principle of Counting If there are \( k \) stages in a process, with \( n_i \) possible outcomes for stage \( i \), then the number of possible outcomes for all \( k \) stages taken together is \( n_1n_2n_3 \cdots n_k \).

gap On a plot, the space that separates clusters of data.

geometric (waiting-time) distribution The distribution of the random variable \( X \) in which \( X \) represents the number of trials needed to get the first success in a series of independent trials, where the probability of a success is the same on each trial.

goodness-of-fit test A chi-square test used to determine whether it is reasonable to assume that a sample came from a population in which, for each category, the proportion of outcomes in the population that fall into that category is equal to some hypothesized proportion.

heteroscedasticity The tendency of points on a scatterplot to fan out at one end, indicating that the relationship varies in strength.

histogram A plot of a quantitative variable that groups cases into rectangles or bars. The height of the bar shows the frequency of measurements within the interval (or bin) covered by the bar.

homogeneous populations Two or more populations that have nearly equal proportions of units in each category of study.

hypothesis test See test of significance.

incorrect response bias A bias resulting from responses that are systematically wrong, such as from intentional lying, inaccurate measurement devices, faulty memories, or misinterpretation of questions.

independent events Events \( A \) and \( B \) for which the probability of event \( A \) happening doesn't depend on whether event \( B \) happens. Events \( A \) and \( B \) are independent if and only if \( P(A|B) = P(A) \) or, equivalently, \( P(B|A) = P(B) \) or, equivalently, \( P(A \text{ and } B) = P(A) \cdot P(B) \).

inference (or inferential statistics) Using results from a random sample to draw conclusions about a population or using results from a randomized experiment to compare treatments.

influential point On a scatterplot, a point that strongly influences the regression equation and correlation. To judge a point's influence, you compare the regression equation and correlation computed first with and then without the point.

interpolation Making a prediction when the value of the explanatory variable, \( x \), falls inside the range of the observed data.

interquartile range, \( IQR \) A measure of spread equal to the distance between the upper and lower quartiles; \( IQR = Q_3 - Q_1 \).

joint frequency The frequency within a particular cell of a two-way table of frequencies for categorical variables.

judgment sample A sample selected using the judgment of an expert to choose units that he or she considers representative of a population.

Law of Large Numbers A theorem that guarantees that the proportion of successes in a random sample
will converge to the population proportion of successes as the sample size increases. In other words, the difference between a sample proportion and a population proportion must get smaller (except in rare instances) as the sample size gets larger, if the sample is randomly selected from that population.

**least squares line**  See **regression line**.

**level**  One of the values or categories making up a factor.

**level of significance, \( \alpha \)**  The maximum \( P \)-value for which the null hypothesis will be rejected.

**line of averages**  See **line of means**.

**line of means (or line of averages)**  Another term for the regression line, if points form an elliptical cloud. In theory, the population regression line contains the means (expected values) of the conditional distribution of \( y \) at each value of \( x \).

**linear shape**  The characteristic of an elliptical cloud of points where the means of the conditional distributions of \( y \) given \( x \) tend to fall along a line.

**lower quartile (or first quartile, \( Q_1 \))**  In a distribution, the value that separates the lower quarter of values from the upper three-quarters of values. The median of the lower half of all the values.

**lurking variable**  A variable other than those being plotted that possibly can cause or help explain the behavior of the pattern on a scatterplot. More generally, a variable that is not included in the analysis but, once identified, could help explain the relationship between the other variables.

**margin of error, \( E \)**  Half the length of a confidence interval; \( E = (\text{critical value}) \cdot \text{(standard error)} \).

**marginal frequency**  The total of the joint frequencies across row categories for a given column or across column categories for a given row of a two-way table of frequencies for categorical variables.

**marginal relative frequency**  The marginal frequency of a two-way table of categorical data divided by the total frequency (number of units represented in the table).

**matched pairs design**  See **randomized paired comparison design**.

**maximum**  The largest value in a data set.

**mean, \( x\)**  A measure of center, often called the average, computed by adding all the values of \( x \) and dividing by the number of values, \( n \). On a plot, the place where you would put a pencil point below the horizontal axis in order to balance the distribution.

**measure of center**  A single-number summary that measures the “center” of a distribution; usually the mean (or average). Median, midrange, mode, and trimmed mean are other measures of center.

**measure of spread (or measure of variability)**  A single-number summary that measures the variability of a distribution. Range, IQR, standard deviation, and variance are measures of spread.

**median**  A measure of center that is the value that divides an ordered set of values into two equal halves. To find it, you list all the values in order and select the middle one or, if the number of values is even, the average of the two middle ones. If there are \( n \) values, the median is at position \( (n + 1) / 2 \). On a plot of a distribution, the median is the value that divides the area between the distribution curve and the \( x \)-axis in half.

**method of least squares**  A general approach to fitting functions to data by minimizing the sum of the squared residuals (or errors).

**midrange**  The midpoint between the minimum and maximum values in a data set, or \( (\text{max} + \text{min}) / 2 \).

**minimum**  The smallest value in a data set.

**mode**  A measure of center that is the value with the highest frequency in a distribution. On a plot of a distribution, it occurs at the highest (maximum) peak.

**modified boxplot**  A graphical display like the basic boxplot except that the whiskers extend only as far as the largest and smallest non-outliers (sometimes called adjacent values) and any outliers appear as individual dots or other symbols.

**Multiplication Rule**  For any two events \( A \) and \( B \), \( P(A \text{ and } B) = P(A) \cdot P(B | A) = P(B) \cdot P(A | B) \). If events \( A \) and \( B \) are independent, then \( P(A \text{ and } B) = P(A) \cdot P(B) \).

**mutually exclusive events**  See **disjoint events**.

**negative trend**  The tendency of a cloud of points to slope downward as you go from left to right, or the tendency of the value of \( y \) to get smaller as the value of \( x \) gets larger.

**nonresponse bias**  A bias that can occur when people selected for the sample do not respond to the survey.

**normal distribution**  A useful probability distribution that has a symmetric bell or mound shape and tails extending infinitely far in both directions.
null hypothesis  The standard or status quo value of a parameter that is assumed to be true in a test of significance until possibly refuted by the data in favor of an alternative hypothesis.

observational study  A study in which the conditions of interest are already built into the units being studied and are not randomly assigned.

one-sided (one-tailed) test of significance   A test in which the P-value is computed from one tail of the sampling distribution. Used when the investigator has an indication of which way any deviation from the standard should go, as reflected in the alternative hypothesis.

outlier  A value that stands apart from the bulk of the data.

parameter  A summary number describing a population or a probability distribution.

percentile  The quantity associated with any specific value in a univariate distribution that gives the percentage of values in the distribution that are equal to or below that specific value. The median is the 50th percentile.

placebo  A nontreatment that mimics the treatment(s) being studied in all essential ways except that it does not involve the crucial component.

placebo effect  The phenomenon that when people believe they are receiving the special treatment, they tend to do better even if they are receiving the placebo.

plot of distribution  (or graphical display or statistical graphic) A graphical display of the distribution of a variable that provides a sense of the distribution's shape, center, and spread.

point estimator  A statistic from a sample that provides a single point (number) as a plausible value of a population parameter.

point of averages   The point \((x, \bar{y})\), where \(\bar{x}\) is the mean of the explanatory variable and \(\bar{y}\) is the mean of the response variable. This point falls on the regression line.

pooled estimate  The weighted average of two statistics estimating the same parameter, with the weights usually determined by the sample sizes or degrees of freedom.

population  The entire set of people or things (units) that you want to know about.

population size  The number of units in the population.

population standard deviation, \(\sigma\)   See standard deviation of a population.

positive trend  The tendency of a cloud of points to slope upward as you go from left to right, or the tendency of the value of \(y\) to get larger as the value of \(x\) gets larger.

power of a test  The probability of rejecting the null hypothesis.

power relationship  A relationship between two variables in which the response variable, \(y\), is proportional to the explanatory variable, \(x\), raised to a power. Mathematically, \(y = ax^b\), where \(a\) and \(b\) are constants.

predicted (or fitted) value  An estimated value of the response variable calculated from the known value of the explanatory variable, \(x\), often by using a regression equation.

prediction error  The difference between the actual value of \(y\) and the value of \(y\) predicted from a regression line. Usually unknown except for the points used to construct the regression line, whose prediction errors are called residuals.

predictor  See explanatory variable.

probability  A number between 0 and 1, inclusive (or between 0% and 100%), that measures how likely it is for a chance event to happen. At one extreme, events that can't happen have probability 0. At the other extreme, events that are certain to happen have probability 1.

probability density  A probability distribution, such as the normal or \(\chi^2\) distribution, where \(x\) is a continuous variable and probabilities are identified as areas under a curve.

probability distribution  See distribution, probability.

probability model (or chance model)  A description that approximates—or simulates—the random behavior of a real situation, often by giving a description of all possible outcomes with an assignment of probabilities.

probability sample  A sample in which each unit in the population has a known probability of ending up in the sample.

protocol  A written statement telling exactly how an experiment is to be designed and conducted.
**P-value** For a test, the probability of seeing a result from a random sample that is as extreme as or more extreme than the one computed from the random sample, if the null hypothesis is true. (Sometimes called the observed significance level.)

**quantitative variable** (or numerical variable)
A variable that takes on numerical values.

**quartiles** Three numbers that divide an ordered set of data values into four groups of equal size.

**questionnaire bias** Bias that arises from how the interviewer asks and words the survey questions.

**random sample** A sample in which individuals are selected by some chance process. Sometimes used synonymously with simple random sample.

**random variable** A variable that takes on numerical values determined by a chance process.

**randomization** (or random assignment) Assigning subjects to different treatment groups using a random procedure.

**randomized block design** An experimental design in which similar units are grouped into blocks and treatments are then randomly assigned to units within each block.

**randomized comparative experiment** An experiment in which two or more treatments are randomly assigned to experimental units for the purpose of making comparisons among treatments.

**randomized paired comparison (matched pairs)** An experimental design in which two different treatments are randomly assigned within pairs of similar units.

**randomized paired comparison (repeated measures)** An experimental design in which each treatment is assigned (in random order) to each unit.

**range** A measure of spread equal to the difference between the maximum and minimum values in a data set.

**rare events** Values or outcomes that lie in the outer 5% of a distribution or in the upper 2.5% and lower 2.5% of a distribution. Compare reasonably likely.

**reasonably likely** Describes values or outcomes that lie in the middle 95% of a distribution. Compare rare events.

**recentering** Adding the same number \( c \) to all the values in a distribution. This procedure doesn't change the shape or spread but slides the entire distribution by the amount \( c \), adding \( c \) to the measures of center.

**rectangular distribution** A distribution in which all values occur equally often.

**regression** The statistical study of the relationship between two (or more) quantitative variables, such as fitting a line to bivariate data. (Can be extended to categorical variables.)

**regression effect** (or regression toward the mean)
On a scatterplot, the difference between the regression line and the major axis of the elliptical cloud.

**regression line** (or least squares line or least squares regression line) The line for which the sum of squared errors (residuals), SSE, is as small as possible.

**regression toward the mean** See regression effect.

**relative frequency** A proportion computed by dividing a frequency by the number of values in the data set.

**relative frequency histogram** A histogram in which the length of each bar shows proportions (or relative frequencies) instead of frequencies.

**repeated measures design** See randomized paired comparison (repeated measures).

**replication** Repetition of the same treatment on different units.

**rescaling** Multiplying all the values in a distribution by the same nonzero number \( d \). This process doesn't change the basic shape but instead stretches or shrinks the distribution, multiplying the IQR and standard deviation by \( |d| \) and multiplying the measures of center by \( d \).

**residual** (or error) For points used to construct the regression line, the difference between the observed value of \( y \) and the predicted value of \( y \), that is, \( y - \hat{y} \).

**residual plot** A scatterplot of residuals, \( y - \hat{y} \), versus predictor values, \( x \), or versus predicted values, \( \hat{y} \). A diagnostic plot used to uncover nonlinear trends in a relationship between two variables.

**resistant to outliers** Describes a summary statistic that does not change very much when an outlier is removed from the data set.

**response variable** The outcome variable used to compare results of different treatments in an experiment or the outcome variable that is predicted by the explanatory variable or variables in regression analysis. Placed on the \( y \)-axis in a regression analysis.
robustness  The comparative insensitivity of a statistical procedure to departure from the assumptions on which the procedure is based.

row marginal frequency  The total of all frequencies across column categories for a particular row of a two-way table of frequencies for categorical variables.

sample  The set of units selected for study from the population.

sample selection bias (or sampling bias or bias due to sampling)  The extent to which a sampling procedure produces samples for which the estimate from the sample is larger or smaller, on average, than the population parameter being estimated.

sample space  A complete list or description of disjoint (mutually exclusive) outcomes of a chance process.

sampling bias  See sample selection bias.

sampling distribution  The distribution of a sample statistic under some prescribed method of probability sampling.

sampling distribution of a sample proportion, \( \hat{p} \)  The theoretical distribution of the sample proportion in repeated random sampling.

sampling distribution of the sample mean, \( \bar{x} \)  The theoretical distribution of the sample mean in repeated random sampling.

sampling frame (or frame)  The listing of units from which the sample is actually selected.

sampling with replacement  In sequential sampling of units from a population, a procedure in which each sampled unit is placed back into the population before the next unit is selected.

sampling without replacement  In sequential sampling of units from a population, a procedure in which each sampled unit is not placed back into the population before the next unit is selected.

scatterplot  A plot that shows the relationship between two quantitative variables, usually with each case represented by a dot.

segmented bar graph (or stacked bar graph)  A plot in which categorical frequencies are stacked on top of one another.

sensitive to outliers  Describes a summary statistic that changes considerably when an outlier is removed from the data set.

shape  One of the characteristics, along with center and spread, that is used to describe distributions. Univariate distributions sometimes have a standard shape such as normal, uniform, or skewed. Bivariate distributions may form an elliptical cloud. Descriptions of shape should consider possible outliers, clusters, and gaps.

simple random sample, SRS  A sample generated from a sampling procedure in which all possible samples of a given fixed size are equally likely.

simulation  A procedure that uses a probability model to imitate a real situation. Often used to compare an actual result with the results that are reasonable to expect from random behavior.

size bias  A type of sample selection bias that gives units with a larger value of the variable a higher chance of being selected.

skewed  Describes distributions that show bunching at one end and a long tail stretching out in the other direction. Often happens because the values “bump up against a wall” and hit either a minimum that values can’t go below or a maximum that values can’t go above.

skewed left  A skewed distribution with a tail that stretches left, toward the smaller values.

skewed right  A skewed distribution with a tail that stretches right, toward the larger values.

slope  For linear relationships, the change in \( y \) (rise) per unit change in \( x \) (run).

split stem  A stem-and-leaf plot in which the leaves for each stem are split onto two or more lines. For example, if the second digit is 0, 1, 2, 3, or 4, it is placed on the first line for that stem. If the second digit is 5, 6, 7, 8, or 9, it is placed on the second line for that stem.

spread  See variability.

stacked bar graph  See segmented bar graph.

standard deviation of a population, \( \sigma \)  A measure of spread equal to the square root of the sum of the squared deviations divided by \( n \). For a probability distribution, it is the square root of the expected squared deviation from the mean.

standard deviation of a sample, \( s \)  A measure of spread equal to the square root of the sum of the squared deviations divided by \( n - 1 \).

standard error  The standard deviation of a sampling distribution.
standard error of the mean, \( \sigma_x \) The standard deviation of the sampling distribution of \( \bar{x} \), or \( \sigma_x/\sqrt{n} \).

standard error of the mean (estimated) The estimated standard deviation of the sampling distribution of \( \bar{x} \), or \( s/\sqrt{n} \).

standard normal distribution A normal distribution with mean 0 and standard deviation 1. The variable along the horizontal axis is called a z-score.

standard units, \( z \) The number of standard deviations a given value lies above or below the mean:
\[
z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}
\]

standardizing Converting to standard units; the two-step process of recentering and rescaling that turns any normal distribution into a standard normal distribution.

statistic A summary number calculated from a sample taken from a population. For example, the sample mean, \( \bar{x} \), and standard deviation, \( s \), are statistics.

statistically significant Describes the situation when the difference between the estimate from the sample and the hypothesized parameter is too big to reasonably be attributed to chance variation.

statistics (or data analysis) The study of the production, summarization, and analysis of data, along with the processes for drawing conclusions from the data.

stem-and-leaf plot (or stemplot) A graphical display with “stems” showing the leftmost digit of the values separated from “leaves” showing the next digit or set of digits.

stemplot See stem-and-leaf plot.

strata (singular, stratum) Subgroups of the population, usually selected for homogeneity or sampling convenience, that cover the entire population. See also stratified random sampling.

stratification A classification of the units in a population into homogeneous subgroups, known as strata, prior to sampling.

stratified random sampling Stratifying the population and then taking a simple random sample from within each stratum.

strength In the context of regression analysis, two variables are said to have a strong relationship if there is little variation around the regression line. If there is a lot of variation around the regression line, the relationship is weak.

sum of squared errors, SSE The sum of the squared residuals: \( \sum (y - \hat{y})^2 \).

summary statistic See statistic.

systematic sampling with random start A sample selected by taking every \( n \)th member of the population, starting at a random spot—for example, having people count off and then picking one of the numbers at random.

table of random digits A string of digits constructed in such a way that each digit, 0 through 9, has probability \( \frac{1}{10} \) of being selected and each digit is selected independently of the previous digits.

t-distribution The distribution, for example, of the statistic below, when the data are a random sample from a normally distributed population:
\[
t = \frac{\bar{x} - \mu}{s/\sqrt{n}}
\]

test of significance (or hypothesis test) A procedure that compares the results from a sample to some predetermined standard in order to decide whether the standard should be rejected.

test statistic Typically, in significance testing, the distance between the estimate from the sample and the hypothesized parameter, measured in standard errors.

third quartile, \( Q_3 \) See upper quartile.

treatment group In an experiment, a group that receives an actual treatment being studied. Compare with control group.

treatments Conditions assigned to different groups of subjects to determine whether subjects respond differently to different conditions.

true regression line (or population regression line) The regression line that would be computed if you had the entire population. Theoretically, the line through the means of the conditional distributions of \( y \) given \( x \). See also line of means.
**t-test**  A test of significance of a population mean (or comparison of means) using the $t$-distribution. See also test of significance and $t$-distribution.

**two-sided (two-tailed) test of significance**  A test in which the $P$-value is computed from both tails of the sampling distribution. Used if the investigator is interested in detecting a change from the standard in either direction.

**two-stage sampling**  A sampling procedure that involves two steps. For example, taking a random sample of clusters and then taking a random sample from each of those clusters.

**two-way table**  A table of frequencies that lists outcomes in the cells formed by the cross-classification of two categorical variables measured on the same units.

**Type I error**  The error made when the null hypothesis is true and you reject it.

**Type II error**  The error made when the null hypothesis is false and you fail to reject it.

**unbiased**  Describes an estimator (statistic) that has an average value in repeated sampling (expected value) equal to the parameter it is estimating.

**uniform distribution**  A distribution whose frequencies are constant across the possible values. Its plot is rectangular.

**unimodal**  Describes a distribution of univariate data with only one well-defined peak.

**units**  Individuals that make up the population from which samples may be selected or to which treatments may be applied.

**univariate data**  Data that involve a single variable per case. A quantitative variable often is displayed on a histogram. A categorical variable often is displayed on a bar chart.

**upper quartile** (or third quartile, $Q_3$)  In a distribution, the value that separates the lower three-quarters of values from the upper quarter of values. The median of the upper half of all the values.

**variability** (or spread)  The degree to which values in a distribution differ. Measures of variability for quantitative variables include the standard deviation, variance, interquartile range, and range.

**variability due to sampling** (or variation in sampling)  A description of how an estimate varies from sample to sample.

**variable**  A characteristic that differs from case to case and defines what is to be measured or classified.

**variance**  A measure of spread equal to the square of the standard deviation.

**voluntary response bias**  The situation in which statistics from samples are not fair estimates of population parameters because the sample data came from volunteers rather than from randomly selected respondents.

**voluntary response sample**  A sample made up of people who volunteer to be in it.

**waiting-time problems**  Problems in which the variable in question is the number of trials you have to wait until the event of interest happens. See also geometric (waiting time) distribution.

**z-score**  See standard normal distribution and standard units.
Brief Answers to Selected Problems

The answers below are not complete solutions but are meant to help you judge whether you are on the right track. If you round computations in intermediate steps or use tables rather than a calculator, your numerical answers might not match those given exactly.

Chapter 1

Section 1.1

P1. Older hourly workers were far more likely to be laid off in Rounds 1–3 than were younger hourly workers.

P2. a. \( \frac{3}{6} \) b. \( \frac{7}{10} \)
   c. A higher proportion of those age 50 and older were laid off than those under age 50 (0.875 versus 0.50).
   d. Hourly workers. The difference in the proportions for salaried workers (0.60 versus 0.375) is smaller than in part c. But note the small number of hourly workers.

P3. hourly, although the patterns are similar

E1. a. \( \frac{14}{27} \) b. \( \frac{4}{9} \)
   c. A higher proportion of those age 40 and older were laid off than those under age 40 (0.52 versus 0.44).
   d. age 50, because the difference in proportions is greater

E3. a. The hourly workers who kept their jobs tended to be younger than the salaried workers who kept their jobs.
   b. no (Salaried workers tended to be older even before layoffs.)

E5. b. The percentage of those laid off who were age 40 or older, by round, were 82%, 89%, 67%, 50%, and 0%. Most layoffs came early, and older workers were hit harder in earlier rounds.

P3. about 37 out of 200, or 0.185

P5. a. 48.6
   b. Write 14 ages on cards. Draw 10 at random and find the average age. Repeat many times and find where 48.6 falls in the distribution.
   c. about 45 out of 200, or 0.225 d. no

Chapter Summary

E15. b. There is no reason to look for an explanation, because there isn't much difference in centers or spreads.

E17. B

E19. a. B d. 13%; no

E21. a. 1001 b. 5, 6, 7, 8, or 9
   c. i. 360 ii. 90 iii. 5 d. \( \frac{455}{1001} \)

Chapter 2

Section 2.1

P1. a. 1
   b. 0.5, 1, and 1.5
   c. 0.5 and 1.5
   d. 15%
   e. 0.05 and 1.95

P2. The number of deaths per month is fairly uniform, with about 190,000–200,000 per month. Summer months have the smallest numbers of deaths, and winter months the largest.
P3. a. A typical SAT math score is roughly 500, give or take about 100 or so.
b. A typical ACT score is about 20, give or take 5 or so.
c. A typical college-age woman is about 65 inches tall, give or take 2.5 inches or so.
d. A typical professional baseball player in the 1910s had a single-season batting average of about .260 or .270, give or take about .040 or so.

P4. The middle 50% of students had GPAs between 2.9 and 3.7, with half above 3.35 and half below.

P5. a. IV  b. II  c. V  d. III  e. I

E1. a. skewed left  b. skewed right  
c. approximately normal  d. skewed right

E5. a. each of the approximately 92 officers; the age at which the officer became a colonel
b. This distribution is skewed left, with no outliers, gaps, or clusters. The middle 50% of the ages are between 50 and 53, with half above 52 and half below.
c. mandatory retirement, age discrimination, an “up or out” rule by which if you haven’t been promoted beyond colonel by your 55th birthday you must retire

e7. approximately normal; too many outliers

E9. a. For example, if a case is a business in the United States and the variable is the number of employees, the distribution will be skewed right. There would be a wall at 1, because that’s the smallest number of employees a business could have (and many businesses have only one employee).
b. For example, if a case is an AP Statistics class and the variable is the percentage of students who did their homework last night, the distribution will be skewed left. There would be a wall at 100%, because in most AP Statistics classes almost all students do their homework.

E11. Births tend to be more frequent in the summer.

E13. a. Within each cluster on either side of the gap from 18,000 to 27,000, the plot is skewed right. If the two clusters are combined (which might not be a good idea), the median is 12,350, with the middle 50% of values between 2,533 and 30,355.

b. Norway and Switzerland; no, they are part of the thinning tail.
c. higher cluster; Eastern Europe, Asia, and the Middle East

d. not when economic conditions of different continents are different

Section 2.2

P6. Quantitative: year of birth, year of hire, and age; categorical: row number, job title, round, and pay category. Month of birth and month of hire are best called “ordered categories.”

P7. The distribution is skewed right, with no obvious gaps or clusters and a wall at 0. The elephant is the only possible outlier. About half the mammals have gestation periods of more than 160 days, and half less. The middle 50% have gestation periods between 63 and 284 days. Large mammals have longer gestation periods.

P8. The average longevity distribution is skewed right, with two possible outliers, while the distribution of maximum longevity is more uniform but has a peak at 20–30 years and a possible outlier. The center and spread of the distribution of maximum longevity are larger.

P9. no

P10. about 0.15; about 30; skewed left, with median between 70 and 75 and the middle 50% between about 60 and 75

P11. See P8.

P12. The cases are the individual males in the labor force age 25 and older. The variable is their educational attainment. The proportion increases through the first three levels with a huge jump at the high school graduation level. Then it decreases except for a spike at bachelor’s degree. The distributions for males and females are similar in shape. Relative frequency bar charts account for the different numbers of males and females.

P13. Westvaco laid off the majority of workers in Rounds 1 and 2.

E15. a. collected by a statistics class; a penny; age of the penny
b. The shape is strongly skewed right, with a wall at 0. The median is 8 years, and the...
spread is quite large, with the middle 50% of
ages falling between 3 and 15 years; however,
it is not unusual to see a penny that is more
than 30 years old.
c. If the same number of pennies is produced
each year and if a penny has the same chance
of going out of circulation each year, then
the height of each column would be a fixed
percentage of the previous height.
E17. D II. A III. C IV. B
E19. a. Twenty percent (or 0.20) of the class got an A.
  b. About 0.47 of the people at the concert
     bought a T-shirt.
E21. b. about 70.2 in.; about 2.8 in.
  c. about 0.93     d. about 0.16
e. The distribution isn't smooth and has a
  domain of only 60.5 to 78.5.
E23. c. Both distributions are slightly skewed right,
with possible outliers on the high end. The
median of both distributions is 12, but the
spread of the distribution of values is larger
for wild mammals.
E25. a. the number of nonpredators that fall into the
categories Domesticated and Wild, and the
total number of nonpredators
b. wild, because the second bar is taller than the
first
c. predator, because the second bar is a larger
fraction of the third bar than is the case for
nonpredators
E27. a. 12     b. two-tenths of 18, or 3.6
  c. The speeds must be estimates for the wild
mammals.

Section 2.3
P14. a. 2.5; 2.5     b. 3; 3     c. 3.5; 3.5
d. 49.5; 49.5     e. 50; 50
P15. about 4 ft 4 in.; about 4 ft
P16. a. Africa: 50; Europe: 80
  b. For Africa, the median is smaller than the
mean because of the right skew. For Europe,
the mean is slightly lower than the median
because of the left skew.
P17. a. 2 and 5; 3     b. 2 and 6; 4
c. 2.5 and 6.5; 4     d. 2.5 and 7.5; 5
P18. a. predators: 12, 7 and 15; nonpredators: 12,
  8 and 15
  b. The distributions are centered at exactly the
same place and have about the same spread,
but the distribution for nonpredators has two
outliers on the high side. Both are essentially
mound-shaped.
P19. a. about 76 million
  b. Half the programs had more than 10 million
     viewers, and half had fewer.
P21. a. min: 1; Q₁: 8; median: 12; Q₃: 15; max: 41
  b. c. −2.5; no outliers on the low end
d. Outliers are elephant at 35 years and
   hippopotamus at 41 years; grizzly bear at
   25 years is the largest non-outlier.
P22. The SD is 3.21.
P23. a. i     b. iii     c. iv     d. vii
e. ii     f. v     g. vi
P24. a. skewed right     b. 3     c. 3.24; 1.89
P25. a. 2.3; 1.84     b. 2 children
  c. positions: 25.5 and 75.5; 2 and 4, with IQR 2;
  1 and 3, with IQR 2
d. The median number of children for 1997 is
  two, one fewer than in 1967; the mean also
went down by about one child per family,
from 3.2 to 2.3. The distributions kept the
same shape and about the same spread.
E29. 10
E31. a. boxplot III     b. boxplot I     c. boxplot II
E33. The back-to-back stemplot is better because there
are only a few values.
E35. laid-off workers: 53.5, 42 and 61; retained
workers: 48, 37 and 55. The median age of the
workers laid off was 5.5 years greater than the
age of those retained, but the distributions have
about the same shape and about the same IQR.
E37. a. II; III     b. II and III
E39. heights of basketball players; heights of all athletes
E41. a. 39.29; 30.425; yes, because it is computed
  using only the maximum and minimum
  b. 10.54
E43. a. 3.11 g     b. 0.043 g     c. yes
E45. a. 72.9
  b. Finding the median of the combined groups
requires having the ordered values.
P26. a. Values tend to be skewed right, because some houses cost a lot more than most houses in a community while very few houses cost a lot less.
   b. $43,964,124.05  c. $4,508.68
P27. a. The distribution of car ages is strongly skewed right.
   b. vehicles proving more durable; people unwilling or unable to buy new cars
P28. a. 4 ft; 3.75 ft; 0.2 ft; 0.25 ft
   b. 50 in.; 47 in.; 2.4 in.; 3 in.
   c. 4 1/3 ft; 4 1/12 ft; 0.2 ft; 0.25 ft
P29. a. 2; 1  b. 12; 1  c. 20; 10  d. 110; 5
   e. −900; 100
P30. a. Outliers occur above −30 + 1.5(21), or 1.5, so Hawaii, at 12, is an outlier.
   b. count 49; mean −41.5; median about −40; SD about 16; min −80; max between −5 and 0; range between 75 and 80; Q1 about −51; Q3 about −30
P31. 425 and 590; about 505; about 165
E47. a. It depends on the purpose of computing a measure of center. Real estate agents usually report the median because it is lower and tells people that half the houses cost more and half less. The tax collector wants the mean price because the mean times the tax rate times the number of houses gives the total amount of taxes collected.
   b. If the reason is to find the total crop in Iowa, use the mean. An individual farmer might use the median to see whether his or her yield was typical.
   c. Survival times usually are strongly skewed right. Telling a patient only the mean survival time would give too optimistic a picture. The smaller median would inform the person that half the people survive for a longer time and half shorter. On the other hand, the mean would help a physician estimate the total number of hours he or she will spend caring for patients with this disease.

E49. a. The scale goes from 36.67 to 58.89.
   b. Variable | N | Mean | Median | StDev
   HighTemp | 50 | 45.61 | 45.56 | 3.72
   Variable | Min | Max | Q1 | Q3
   HighTemp | 37.78 | 56.67 | 43.33 | 47.78
   c. yes
E51. Show that the mean of \((x_1 + c) + (x_2 + c) + (x_3 + c) + (x_4 + c) + (x_5 + c)\) is equal to the original mean, \(\overline{x}\) plus \(c\).
E53. a. .23¢ or .24¢  b. about 10 and about 70; 60  c. skewed right  d. no
E55. about 32 mi/h or, if rounded down, 30 mi/h

P32. a. 1.29%  b. 4.75%  c. 34.46%  d. 78.81%
P33. a. \(\bar{x} = 0.47\)  b. \(\bar{x} = 0.23\)  c. 1.13  d. 1.555
P34. a. 85.58%  b. 99.74%
P35. a. −1.645 to 1.645  b. −1.96 to 1.96
P36. a. cancer, because it is 1.35 SDs below the mean, compared to 0.808 SD for heart disease
   b. cancer, because it is 1.23 SDs above the mean, compared to 1.097 SDs for heart disease
   c. The death rate for heart disease in Colorado (1.83 SDs below the mean) is more extreme than the death rate for cancer in Hawaii (1.29 SDs below the mean).
P37. a. about 24.2%  b. about 63.8 in.
P38. a. about 152 to 324  b. about 186 to 290
P39. A. outside both  B. not outside either  C. not outside either  D. not outside either
E59. a. 84.13%; 99.43%  b. 15.87%; 0.57%
   c. 93.32%  d. 68.27% (68.26% using Table A)
E61. a. 2  b. 1  c. 1.5  d. 3  e. −1  f. −2.5
E63. a. i. 0.6340 (calculator: 0.6319)
   ii. 0.0392 (calculator: 0.0395)
   iii. 0.3085 (calculator: 0.3101)
   b. about 287 to 723
E65. 76%; 87%
E67.  a. 0.1587  b. 8.16 (or 8.17)  
c. 10.02  d. 8 (or 7.89)

E69.  68%; 95%; 16%; 84%; 97.5%; 2.5%

E71.  a. 0.2177 (or 0.2183)  
b. 1,203,800 (or 1,207,650)  
c. 73.56 in.

E73.  a. skewed right  
b. 0.1151  
c. If all values in the distribution must be positive and if two standard deviations below the mean is less than 0, the distribution isn't approximately normal.

Chapter Summary

E75.  b. min: 0; Q₁: 2; median: 10; Q₃: 35; max: 232  
c. Florida and Texas  
e. Both show the strong skewness and outliers. In the stemplot, you can see that half the states have fewer than ten tornadoes. The cleanliness of the boxplot makes it clear how much of an outlier Texas actually is. However, you can't see from the boxplot that many states have at most one tornado. Because the stemplot is easy to read while showing the values, it is reasonable to select it as the more informative plot.
   f. The distribution is strongly skewed right, with two outliers and a wall at 0. The median number of tornadoes is 10, with the middle 50% of states having between 2 and 35 tornadoes.

E77.  a. scores below 457.5 or above 797.5  
b. probably skewed right

E79.  a. Region 1 is Africa; Region 2 is the Middle East; Region 3 is Europe.  
b. Region 1 is A; Region 2 is B; Region 3 is C.

E81.  No; for example, values {2, 2, 4, 6, 8, 8} are symmetric with mean and median 5. Only two of the six values, or about 33%, are within one SD, 2.76, of the mean, 5.

E83.  a. Half the cities have fewer than 47 pedestrian deaths per year, and half have more.  
b. The three outliers—Chicago, Los Angeles, and New York—are the three most populous cities in the United States.  
c. A stemplot reveals the three outliers and the right skew and retains the original values.  
d. If the data were presented in rates per 100,000 population, the three cities in part b might not be outliers.

E85.  2.98; 1.33

E87.  An example is {1, 1, 1, 1, 2, 2, 10}.

E89.  a. i. 6.325 vs. 6.667  ii. 2.000 vs. 2.010  iii. 0.632 vs. 0.633  
b. No; as n gets larger, the difference between s and \( \sigma \) approaches 0.

E91.  a. The state with the lowest average income in dollars in 1980 had an average income of $7,007.  
b. There is at least one outlier on the high end for both years. There are no outliers on the low end for either year.
   c. no

E93.  126.65 mg/dL and 225.35 mg/dL.

E95.  a. about .260; about .040.  
b. Both distributions are approximately normal in shape. The distribution for the American League has a higher mean (by about .010) and less spread.
   c. about .300

Chapter 3

Section 3.1

P1.  b. linear, positive, strong  
c. yes; no  
d. maturation during the early years of life

P2.  a. Delta; Northwest  
b. upper left; United and America West  
c. false  d. negative; weak
   e. no, because these are the largest carriers in the United States; no; yes

E1.  a. positive, strong, linear  
b. negative, strong, linear  
c. positive, moderate, linear  
d. negative, moderate, linear  
e. positive, strong, linear  
f. negative, strong, curved  
g. negative, strong, curved; more variability among values of y for smaller values of x than for larger values of x  
h. positive, strong, curved

E3.  A. II  B. IV  C. III  D. I
E5.  a. i. A, B, and C; none ii. A, B, and C; D iii. A and C
b. Positive: A and C; negative: B; D has little trend.
c. A and C; D
d. no, because these are highly rated universities
e. Graduation rates may increase as SAT scores increase because better-prepared students are more successful in college. Alumni giving rates may increase as graduation rates increase because there are more happy alumni.

E7.  a. Cases are the individual employees at the time of the layoffs; variables are the age at hire and the year of hire. There is a weak positive association with heteroscedasticity. There are no points in the upper left because these employees have reached retirement age.

b. No; you need a plot of the age at hire of all people hired, not just those who remained at the time of layoffs.
c. People hired earliest were more likely to be laid off. Perhaps they had obsolete job skills.

Section 3.2

P3.  Each day the eraser tended to lose around 0.0135 g.
P4.  a. about 0.8
b. If one student has a hand length that is 1 in. longer than that of another student, the first student’s hand tends to be 0.8 in. wider.
c. \[\text{hand width} = 1.7 + 0.8 \times \text{hand length}\]
d. The students in the lower cluster did not spread their fingers. The slope and intercept would be larger.
P5.  a. student/faculty ratio; alumni giving rate
b. A horizontal run of 5 corresponds to a drop of about 10 percentage points.
c. no, because no university has a student/ faculty ratio of 0 (i.e., no students)
d. about 23%; large e. about 1 or 2; about 16 f. \(-11\) g. positive
P6.  b. \[\hat{y} = 279.75 + 2.75x\]
c. For every additional gram of fat, a pizza tends to have 2.75 more calories. Five ounces of pizza with no fat is predicted to have 279.75 calories.
d. \[279.75 + 2.75(11) = 310\]
e. \(0.5 + (-1.0) + 0.5 = 0\)
P7.  % on time = \(87.2 - 2.15\) mishandled baggage;
4.08, 2.31, 0.71, 6.51, \(-1.56\), \(-0.66\), \(-0.11\), 0.18, \(-2.99\), \(-8.47\)
P8.  yes; in row “Error” and column “Sum of Squares”
E9.  a. I—E; II—C; III—A; IV—D; V—B
b. I—A; II—E; III—B; IV—D; V—C
E11.  a. about 2.5 in./yr
b. Boys tend to grow a median of about 2.5 in. per year from age 2 to 14.
c. \[height = 31.5 + 2.5 \times \text{age}\]
d. An average newborn is 31.5 in. long. No; this is too long.
E13.  a. The points lie on a line.  b. 0

b. For every 1 mi/h increase in speed, the reaction distance is an additional 1.1 ft.
c. \[\hat{y} = 1.1x,\] where \(\hat{y}\) is predicted reaction distance in feet and \(x\) is speed in miles per hour
d. 60.5 ft; 82.5 ft
e. The equation would be \(\hat{y} = 1.47x\).
P15.  a. arsenic concentration in the well water; concentration of arsenic in the toenails of people who use the well water
b. moderate, positive, linear; cluster in lower left
c. about 0.3 ppm d. about 0.4 ppm
e. exceeded by seven wells
E17.  a. \[\hat{y} = 38 - 6x,\] where \(\hat{y}\) is the predicted number of days in Detroit with AQI greater than 100 and \(x\) is the number of years after 2000
b. The number of days with AQI greater than 100 in Detroit tended to decrease by 6 per year on average.
c. 2002; 2 d. 6 e. \(-1 + 2 + (-1) = 0\)
f. For \(y = 40 - 6x, \text{SSE} = 18\). So the first line fit better.
g. \(y = 37 - 6x; 25; 0, 0, 3; 9\)
h. Yes; for both of the other lines, all points are on or to the same side of the line.
E19.  a. \[height = 31.57 + 2.43 \times \text{age}\]
b. positive; 1.20 (or 1.21 with no rounding)
c. 51 = 31.57 + 2.43 \times 8 (except for rounding error)
d. almost the same
E21. a. $\text{height} = 31.5989 + 2.47418\text{ age}$ b. 2.4; yes
E23. $-1.33, 0.69, -1.99$; possible curvature
E25. a. Yes; $\hat{y} = 195 + 10.05x$, where $\hat{y}$ is the predicted number of calories and $x$ is the number of grams of fat; for every additional gram of fat, a pizza has about 10 more calories.
b. Yes; $\hat{y} = 10.7 + 2.41x$, where $\hat{y}$ is the predicted number of grams of fat and $x$ is the cost; for every additional dollar in cost, the number of grams of fat tends to increase by 2.41 g.
c. no
d. Calories has a moderately strong positive association with fat, which makes sense because fat has a lot of calories. Fat has a weak positive association with cost. There appears to be no association between cost and calories.

Section 3.3

P9. a. $-0.5$ b. $0.5$ c. $0.95$
   d. $0$ e. $-0.95$
P10. b. $0.908$
P11. a. $-1$ b. $0.5$ c. $-0.5$ d. $-1$
   e. $1$ f. $-0.5$ g. $0.5$ h. $-0.5$
P12. $0.908$; all but one are positive, resulting in a positive correlation.
P13. a. positive b. about $(4.8, 5)$
   c. Quadrants I and III; 20
   d. Quadrants II and IV; 7
P14. no for the top plot because of the curvature; yes for the bottom plot
P15. a. $0.650$
   b. Exam 2 $= 48.94 + 0.368$ Exam 1; 78.38
   c. Exam 1 $= -14.1 + 1.149$ Exam 2
P16. a. the city’s population
   b. Divide each number by the population of the city to get the number of fast-food franchises per person and the proportion of the people who get stomach cancer.
P17. Parents tend to give higher allowances to older children, and vocabulary is larger for older children than for younger children.
P18. that people are too busy watching television to have babies; how affluent the people are
P19. a. $-0.747$
   b. The value of $r^2$ means that 55.8% of the state-to-state variability in the percentage of families living in poverty can be “explained” by the percentage of adults who are high school graduates. In other words, there is 55.8% less variability in the differences between $y$ and $\hat{y}$ than between $y$ and $\bar{y}$. So, by knowing the high school graduation percentage for a state and using the regression line, you tend to do a better job of predicting $y$ than if you used just $\bar{y}$ as the prediction for that state.
   c. no
d. percentage of high school graduates; percentage of families living in poverty; percentage of families living in poverty per percent of high school graduates; no units
P20. yes; yes
P21. Yes; an ellipse around the cloud of points will have a major axis that is steeper than the regression line. Also, for the vertical strip containing exam 1 scores above 95, the mean exam 2 score is about 93. For exam 1 scores less than 70, the mean exam 2 score is about 76.
P27. a. 0.66 b. 0.25 c. 0.06 d. 0.40
   e. 0.85 f. 0.52 g. 0.90 h. 0.74
P29. a. 0.707 b. 0.707
P31. a. about 0.95
   b. All but three of the points lie in Quadrants I and III (based on the “origin” $(\bar{x}, \bar{y})$), where each $z_x \cdot z_y$ is positive.
   c. The point in the lower-left corner of Quadrant III is the most extreme in both $x$ and $y$, so $z_x \cdot z_y$ will be largest.
   d. For the point just below $(\bar{x}, \bar{y})$, $z_x$ and $z_y$ are both near 0, so $z_x \cdot z_y$ will be quite small.
P33. a. no b. yes
P35. a. yes b. $s_x = 25$ c. 0.081 d. 0.0183
P37. a. overall size of the animal
   b. Inflation: All costs have gone up over the years.
   c. Time: Stock prices generally go up due to inflation; also the Internet is new technology, so the number of Internet sites is increasing.
E39.  

a. \( r = 0.9030 \)  
b. \( -0.0523 \)  

c. Yes; hot weather causes people to want to eat something cold.  
d. degrees Fahrenheit; pints per person; pints per person per degree Fahrenheit; no units  
e. \( \frac{SS}{DF} \)

E41.  Scoring exceptionally well involves some luck—in the questions asked, feeling well, having no distractions, and so on. It’s unlikely that this combination will happen again on the next test for the same student.

Section 3.4

P22.  

a. There is little pattern except for one outlier in the upper right.  
b. \( \hat{y} = -680 + 2.85x; r = 0.694 \)  
c. \( \hat{y} = 1350 - 2.14x; r = 0.501. \) Now the slope is negative and the correlation is low. Titanic has a huge influence.

P23.  

a. Not well; the estimates were low.  
b. (180, 350)  
c. With the point: \( \text{actual} = 12.23 + 1.92 \text{estimate} \) and \( r = 0.975. \) Without the point: \( \text{actual} = -27.10 + 3.67 \text{estimate} \) and \( r = 0.921. \) This point pulls the right end of the regression line down, decreasing the slope and increasing the correlation.

P24.  

b. Residuals are \(-0.5, 0.5, 0, \) and 0.  
d. A residual plot eliminates the tilt in the scatterplot so that the residuals can be seen as deviations above and below a horizontal line. Here the symmetry of the residuals shows up better on the residual plot.

P25.  

a. A—IV; B—II; C—I; D—III  
b. i. opens upward, as in II  
ii. fans out or in, as in I  
iii. opens downward  
iv. V-shaped, as in III  
c. plot D; residual plot

E43.  

a. no; not elliptical and has an influential point  
b. \( \hat{y} = -161.90 + 0.954x; r = 0.49 \)  
c. With Antarctica removed, the slope of the regression line changes from positive (0.954) to negative (−1.869) and the correlation becomes negative, \( r = -0.45. \) Without a plot, you can't see that there is little relationship.

E45.  

a. \( \hat{y} = -1.63 + 0.745x; \) residuals: \( -0.567, -4.331, -0.096, -0.803, -1.529, 1.159, 7.178, -1.841, 1.433, -0.605 \)  
b. more variability in the middle than at either end, partly because of more cases in the middle  
c. (8, 0); (10, 13)

E47.  

b. Pizza Hut's Stuffed Crust; Pizza Hut's Pan Pizza; no (make a boxplot)  
c. Removing Domino's Deep Dish increases the slope from about 14.9 calories per gram of fat to around 18 calories and decreases the correlation from 0.908 to 0.893.

E49.  

A—I; B—IV; C—III; D—II; C and D

E51.  

a. residuals: \(-66.67, 133.33, -66.67 \)  
b. only the horizontal scale

E53.  To start, the predicted weight of a person whose height is 64 in. is about 145 + 1.2, or 146.2 lb. The slope of the regression line must be about \( \frac{187 - 145}{76 - 64} \), or 3.5.

Section 3.5

P26.  

b. \( \ln y = 5.22 - 0.435x; 1 - e^{-0.435} \approx 0.35, \) or 35%, per time period  
d. some curvature, indicating a death rate of more than 0.35 in the early rolls and less than 0.35 in the later rolls

P27.  

a. log transformation  
b. \( \ln(\text{pop}) = -54.9342 + 0.03583 \text{ year}; \) 3.6% per year  
c. Florida grew less rapidly than the model predicts until about 1845, then grew more rapidly than predicted, then less, then more. There was a big jump in growth between 1950 and 1960 and a big drop in 2000.

P28.  

(2, 3), (1, 2), (0, 1), (−1, 0); slope 1 and \( y\)-intercept 1

P29.  

a. (6, 3), (4, 2), (2, 1), (0, 0); slope 0.5 and \( y\)-intercept 0  
b. (5, −4), (6, −2), (8, 2); slope 2 and \( y\)-intercept −14

P30.  for P28: \( y = 10(10)^x; \) for P29 part a: \( y = 1(10^{0.5})^x = 3.16^x; \) for P29 part b: \( y = 10^{-14} \cdot 100^x \)
P31. \[ \text{flight length} = 4.807(1.012)^{speed} \]

P32. Taking the natural log of the consumption gives \[ \ln(\text{consumption}) = 0.39 + 0.143 \cdot \text{trips}, \] with \( r = 0.69 \). However, the transformed points don't form an elliptical cloud. Some fishermen's families eat essentially no fish, even if the person fishes as many as 11 times a month.

P33. about 54.9 lb (or 54.8 lb with no rounding)

P34. a. There is a strong positive curved relationship, with the rate of change in velocity decreasing as the depth increases.
   
   b. \[ \ln(\text{velocity}) = 0.146 + 0.175 \ln(\text{depth}) \]

P35. a. \( \frac{1}{3} \) (cube root)  
   b. \( -1 \) (reciprocal)  
   c. 2 (square)

P36. a. 3 (cube)  
   b. \( -1 \) (reciprocal)  
   c. \( \frac{1}{2} \) (square root)  
   c. \( \frac{1}{2} \) (square root)

P37. a. \( \frac{1}{2} \) (square root)  
   b. \( \frac{1}{3} \) (cube root)  
   c. \( \frac{1}{2} \) (square root)

P38. diameter squared versus age; diameter versus square root of age

P39. a. less than 1  
   b. nearly linear  
   c. \[ \log(\text{brain}) = 0.908 + 0.76 \log(\text{body}) \]
   or \[ \text{brain} = 8.10 \text{body}^{0.76}; \] yes

E55. no

E57. A log-log transformation does not remove all the curvature. If you split the ages into two groups, between 8 and 9, and fit a linear equation to each group separately, the residuals show less curvature and are smaller.

E59. a. strong positive relationship with some curvature  
   b. A line is not a bad fit here and would predict reasonably well for parties of 16 or fewer chimps because the residuals are small. Because of the curvature, it probably would not predict as well for parties much larger than 16 chimps.  
   c. A log-log transformation works pretty well. The model would be \[ \ln(\text{percent}) = 0.524 \ln(\text{chimps}) + 2.9575 \text{ or percent} = 19.25 \text{chimps}^{0.524}. \]  
   d. The residual plot shows a random scatter, which is good; however, there is more spread for the smaller hunting parties than for the larger ones, so a transformation that reduces this would be better.

E61. There is essentially no relationship between the day these passengers bought their tickets and the price they paid, with the exception of five passengers who bought their tickets within 9 days of the flight and paid more than double what any other passenger paid.

E63. a. negative and curved  
   b. \[ \log(\text{GNP}) = 1.87 - 0.0674 \text{ birthrate} \]
   or \[ \text{GNP} = 74.13 \cdot 0.856^{\text{birthrate}}; \log(\text{GNP}) \text{ decreases, on average, 0.0674 unit for every 1 unit increase in birthrate. A country with no births is predicted to have a log(\text{GNP}) of 1.87 (which doesn't make sense).} \]

E65. a. There is an increase in \( CO_2 \) over the years, with upward curvature.  
   b. \( CO_2 \) increased at a lower rate than the overall average from 1967 to about 1994 and at a higher rate than the overall average during the very beginning and the very end.  
   c. Use different lines, one line for 1959 to about 1976 and the other for 1977 to 2002. Or recognize that, while an exponential model has an asymptote at 0, the \( CO_2 \) level in the atmosphere was never near 0. Preindustrial levels of \( CO_2 \) were around 250 ppm. Adjust for this by taking the natural log of \( (\text{CO}_2 \text{ level} - 250) \).
   d. The linear model for years after 1976 gives an average increase of about 1.57 ppm \( CO_2 \) per year. Using the exponential model with an asymptote at 250 ppm gives a growth rate of about 1.5% per year.

Chapter Summary

E67. a. Arm span from height: \( \hat{y} = -5.81 + 1.03x \); kneeling height from height: \( \hat{y} = 2.19 + 0.73x \); hand length from height: \( \hat{y} = -2.97 + 0.12x \);  
   yes, the slopes are about what he predicted (1, 0.75, and \( \frac{1}{3} \)), and the \( y \)-intercepts are close to 0 in each case.  
   b. If one student is 1 cm taller than another, the arm span tends to be 1.03 cm larger. If one student is 1 cm taller than another, kneeling height tends to be 0.73 cm larger. If one student is 1 cm taller than another, hand length tends to be 0.12 cm longer.
c. \textit{arm span and height}: 0.992 (strongest), \textit{kneeling height and height}: 0.989, \textit{hand length} and \textit{height}: 0.961 (weakest)

E69. a. Yes; (52, 83) lies away from the general pattern on the scatterplot; the residual plot shows that the student scored much higher than expected on the second exam.

b. The slope increases from 0.430 to 0.540, and the correlation increases from 0.756 to 0.814.

c. yes

d. Yes; the student who scored lowest on Exam 1 did much better on Exam 2, and the highest scorer on Exam 1 was not the highest scorer on Exam 2.

E71. a. Match each value with itself.

b. You can get a correlation of 0.950.

c. You can get a correlation of $-0.1$.

d. Match the biggest with the smallest, the next biggest with the next smallest, and so on.

E73. a. true b. true c. false d. true

E75. a. Public universities that have the highest in-state tuition also tend to be the universities with the highest out-of-state tuition, and public universities that have the lowest in-state tuition also tend to be the universities with the lowest out-of-state tuition. This relationship is quite strong.

b. no; yes c. no; yes

E77. C, B, A

E79. For example, stocks that do best in one quarter may not be the ones that do best in the next quarter.

E81. a. very strong linear relationship; one influential point

b. very strong linear relationship; one influential point

c. Yes; in general, the more police, the higher the rate of violent crime; a log-log transformation straightens these data.

E83. a. moderate positive correlation of 0.66; cause and effect, because if teachers are paid more the cost per pupil has to go up (unless class sizes are increased proportionally)

b. moderate positive correlation of 0.577; about the same; yes, because the number of pupils in the state is pretty much proportional to the number of people in the state

c. no

\textbf{Chapter 4}

\textbf{Section 4.1}

P1. a. all the households in your community; individual households

b. quicker, cheaper, and easier

c. getting a response from each household


P3. a. size bias

b. The bias isn't obvious, but it is likely there was some because of the judgment sampling.

c. voluntary response bias

d. size bias

e. voluntary response bias

P4. too high; convenience sampling

P5. question I

P6. incorrect response bias

E1. a. too high

b. Adults are more likely to be home.

c. sampling bias

E3. convenience sample; sample selection bias; probably too high

E5. too high; too high

E7. a. voluntary response

b. No; people with stronger feelings are more likely to respond.

c. quite a bit less than 92%

E9. Too high; the more children, the more likely a family is to be in the sample.

E11. ABC News poll

\textbf{Section 4.2}

P7. a. no

b. Although this produces a random selection of students, it does not produce a simple random sample of a fixed sample size.

c. No; two students sitting in different rows cannot both be in the sample.

d. yes
e. No; a group of six girls cannot all be in the sample.
f. No; two students with last names starting with different letters cannot both be in the sample.
P8. a. 516; 384; 300  b. age
P9. 0.79
P10. a. Choose a random start between persons 1 and 20, and take every 20th person.
    b. Choose a random start between persons 1 and 5, and take every 5th person.
P11. Choose a random start between 1 and 17, and take every 17th person.
P12. a. Make an estimate of the number of mortgages and divide by the sample size desired. Suppose you get \( k \). Then pick an integer at random from 1 through \( k \), say you get \( j \). The sample consists of the mortgage at position \( j \) in chronological order, and every \( k \)-th mortgage after that.
b. Pick a random sample of dates. The sample consists of all mortgages assumed on those dates.
c. Start as in part b, then take a random sample of mortgages from within each cluster.
P13. a. Use pages as clusters.
    b. Take an SRS of pages. Then take an SRS of lines from each of those pages.
    c. Take an SRS of characters from each line selected in part b.
E15. a. stratified random sample with strata of grocery store owners and restaurant owners
E17. Stratification by gender is likely to be the best strategy.
E19. Consider the farms as the five strata, and take a random sample of, say, 10 acres from each farm.
E21. systematic sampling with random start in both cases

Section 4.3

P14. a. children who live near a major power line and the matching children; living near a major power line or not; whether the child gets leukemia
    b. no, because the children are not randomly assigned to the two conditions
    c. For example, major power lines often are near a major highway (and hence in a polluted area).
P15. a. Randomly select four test sites for each type of glass. Measure the amount of energy lost at each site, and compare the results from the three types of glass.
    b. the twelve site-generator combinations; the type of glass
    c. Randomly assign each treatment to two generators at each site.
P16. a. The lurking variable is the person’s age.
    b. I causes II.  c. II causes I.
P17. | took course? | high | low |
    |-----------------|-----|-----|
    | yes            | higher SAT | no evidence |
    | no             | no evidence | lower SAT |
P18. a. Brightness of the room and type of music; for brightness, the levels are low, medium, and high; for type of music, the levels are pop, classical, and jazz.
    b. Possibilities are heart rate, blood pressure, and self-description of anxiety level.
P19. a. No; the pipe and cigar smokers may be older.
    b. observational study
    c. smoking behavior; nonsmoking, cigarette smoking, and pipe or cigar smoking; number of deaths per 1000 men per year
P20. Older men have a higher death rate, and the pipe and cigar smokers are older; the new factor is age.
P21. a. observational
    b. legal age for driving; the age groups; highway death rate by state
    c. for example, driver education, because states with higher age limits may generally be more restrictive and require more training
P22. a. green
    b. The greater distances traveled by the green bears is due to confounding of launch order with bear color. As students got more practice, they were able to launch the bear farther.
    c. Later launches tend to have greater distances.
    d. Randomize the order in which the bears of different colors are launched.
P23. adults who died of nonrespiratory causes; yes; no; no
P24. a. No; students might have used other clues.  
   b. group with the magnets  
   c. whether the students were assigned randomly to the treatments  
   d. The second design is better than the first.  
   e. no; no  

P25. the two textbooks; the ten classes; five for each treatment  

P26. Carnation plants (if in separate containers); put the plants in place and then randomly assign the new product to half and leave the others growing under standard conditions. There must be at least two plants in each group, and preferably many more. The plants receiving the standard treatment can be used as a control.  

P27. a. brand of paper towel, with levels Brand A and Brand B, and wetness, with levels dry and wet; number of pennies a towel can hold before breaking  
   b. Randomize both the assignment of wet or dry to five towels from each brand and the order of testing; the experimental units are the 20 time-towel combinations for which the tests will be performed.  
   c. experiment  

E23. observational study  

E25. a. yes  
   b. All ten must be dug up so that the variables shade and being dug up are not confounded.  

E27. a. dormitories  
   b. 20  
   c. experiment  

E29. different population sizes; climate is confounded with proportion of older people in the state; observational study  

E31. not unless the subjects are in random order to begin with  

Section 4.4  

P28. a. Some dorms are less healthy than others (too stuffy or contaminated); some may be for athletes or others who tend to be healthier.  
   b. to equalize variability between the treatment groups  

P29. a. randomized paired comparison (repeated measures)  
   b. randomized paired comparison with matched pairs, because the drug might not clear out of the bloodstream in the time allowed between treatments  

P30. randomized paired comparison with repeated measures  

P31. a. There is a lot of variability in how well students memorize, so use a randomized paired comparison design with either matched pairs or repeated measures. The response variable could be the number of words on a list that are remembered.  
   b. Bigger people tend to eat a lot more soup than do smaller people. Blocking could be done by estimating a person’s weight. The response variable is the amount of soup eaten.  

P32. a. No; no; you can’t tell which dots represent measurements for the same patient.  
   b. yes; yes  
   c. Subtract the number of flicks per minute for Treatment C from that for Treatment A.  
   d. Randomized paired comparison with repeated measures; there’s a lot of subject-to-subject variability, as shown in Display 4.20.  

E33. completely randomized; high school students who want to take the special course; course-taking assignment; take course or don’t take course; SAT score; no blocks  

E35. two weeks of a patient’s time; low phenylalanine diet or regular diet; one week of a patient’s time; paired comparison (repeated measures) with, we hope, randomization  

E37. block design; rate of finger tapping; one day of a subject’s time; caffeine, theobromine, or placebo; three days of a subject’s time  

E39. completely at random; quantity eaten (relative to body weight); hornworm; diet of regular food or diet of 80% cellulose; no blocks  

Chapter Summary  

E41. a. sample, because the measurement process is destructive  
   b. census, because the population is small and the information is easy to get  
   c. sample, because a census would be too costly and time-consuming
E43.  
a. probably reasonably representative  
b. not representative, because too young  
c. probably reasonably representative  
e. not representative, because blood pressure tends to increase with age

E45. estimate of earnings too high

E47. owners who sold their 2000 model, who might be the owners unhappy with high repair bills

E49. Your estimate will be too large because your net allows tiny fish to escape.

E51. *New York Times* readers tend to have higher incomes and more years of education.

E53. Choose an SRS of states; choose an SRS of congressional districts from each state; choose an SRS of precincts from each congressional district; choose an SRS of voters from each precinct.

E55.  
a. observational study  
b. The researcher can’t tell whether it was the diet or some other difference in lifestyles that accounts for the result.  
c. more physical activity and less stress in Greece

E57.  
a. experiment; factors: presence of mother, with levels present and absent, and presence of siblings, with levels present and absent; difference in the mouse’s weight; total number of baby mice  
b. survey; the amounts of the ten different nutrients; the number of species of plants; 4  
c. observational study; the different types of fruit; float or not; one observed unit for each type of fruit

E59.  
a. location; urban or rural; one pair of twins  
b. observational study  
c. within the rural twins; yes  
d. genetically identical; variability caused by genetic differences

E61. Form four blocks of spaces with similar locations in the gym: 1 and 4, 2 and 3, 5 and 8, 6 and 7. Randomly assign two bikes within each block.

Chapter 5
Section 5.1

P1.  
a. \[ P(0) = \frac{1}{16} ; P(1) = \frac{4}{16} ; P(2) = \frac{6}{16} ; P(3) = \frac{4}{16} , \]
\[ P(4) = \frac{1}{16} \]

b. \[ \frac{1}{16} \]

c. Probably not; if no one can tell the difference, there is 1 chance in 16 that all four will select correctly.

P2.  
a. \[ \frac{13}{27} , \text{ or about } 0.48 \]

b. For example, for a temperature of 20°F, the probability is \[ \frac{2}{27} , \text{ or about } 0.074. \]

c. too warm

P4.  
a. 28, 35; 28, 41; 28, 47; 35, 41; 35, 47; 35, 55; 41, 47; 41, 55; 47, 55  
b. yes  
c. \[ \frac{1}{10} \]  
d. \[ \frac{6}{10} \]

P5.  
a. yes  
b. no  
c. no  
d. no

P6.  
a. yes  
b. no

P7.  
a. tails; tails; heads; tails; tails  
b. 0.44

P8.  
b. 6  
c. probably not

P9.  
a. 21  
b. \[ \frac{1}{21} \]

E1.  
a. There are 32 outcomes.  
b. \[ P(0) = \frac{1}{32} ; P(1) = \frac{5}{32} ; P(2) = \frac{10}{32} ; P(3) = \frac{10}{32} ; P(4) = \frac{5}{32} ; P(5) = \frac{1}{32} \]

E3.  
a. \[ \frac{30}{36} \]  
b. \[ \frac{4}{36} \]  
c. \[ \frac{8}{36} \]  
d. \[ \frac{6}{36} \]  
e. \[ \frac{11}{36} \]

f. \[ \frac{1}{36} \]  
g. \[ \frac{9}{36} \]  
h. \[ \frac{3}{36} \]  
i. \[ \frac{2}{36} \]

E5.  
a. disjoint and complete; \[ \frac{9}{16} , \frac{6}{16} , \frac{1}{16} \]  
b. not disjoint; complete  
c. disjoint; not complete  
d. disjoint; not complete  
e. disjoint; not complete

E7.  
a. no; no; yes; no  
b. 6  
c. \[ \frac{6}{50} \]  

E9.  
a. 20  
c. 40

E11.  
a. yes  
b. They get the white marble on the same draw; \[ \frac{1}{81} . \]  
c. Add more non-white marbles to the bags.

E13.  
a. \[ 64, \frac{1}{64} \]  
b. \[ 46,656, \frac{1}{46,656} \]  
c. \[ 2^{1200} , \text{ no, because outcomes aren’t equally likely} \]
Section 5.2

P10. a. Assign a digit from 1 to 8 to each worker. Start at a random place in Table D and look at the next three digits. Those digits will represent the workers selected to be laid off. Ignore 9, 0, and any digit that repeats, and select another digit.

b. Assign a pair (or pairs) of digits to each worker.

P11. a. Each teaspoon has probability 0.80 of disappearing. Whether each spoon disappears is independent of whether other spoons disappear.

b. Assign digits 1–8 to disappeared and 9 and 0 to did not disappear. Look at ten digits, allowing repeats, and record how many were 1–8.

c. with the assignments given in part b: disappeared, did not disappear, disappeared, did not disappear, disappeared, did not disappear, disappeared, did not disappear, disappeared, did not disappear

d. The estimated probability that all ten spoons disappear within 5 months is 0.1034.

P12. a. Each possible set of eight catastrophic accidents is equally likely to be in the sample.

b. Assign a triple (or triples) of digits to each accident.

d. The estimated probability that at least half of a random sample of eight catastrophic accidents are from cheerleading is about 0.72.

P13. a. The probability that a given team wins a particular game is 50%, no matter what the results of previous games are.

b. Assign digits 1–5 to team A winning and digits 6–9 and 0 to team B winning. Select digits, allowing repeats, until one of the teams has four wins. Record the total number of digits needed.

d. The estimated probability that a World Series of two evenly matched teams will go seven games is 0.31.

E15. a. The probability that a randomly selected high school girl reports that she rarely or never wears a seat belt is 10%. The girls are selected randomly and independently.

b. Assign the digit 0 to reports she rarely or never wears a seat belt and the other digits to does not report that she rarely or never wears a seat belt. Select four digits, allowing repeats. Record the number of digits that are 0.

d. The estimated probability that a random sample of four girls contains no more than one who says she rarely or never wears a seat belt is about 0.95.

E17. a. Each question has a 25% chance of being answered correctly, regardless of whether other questions are answered correctly.

b. Assign pairs of digits 01–25 to a correct response and pairs 26–99 and 00 to an incorrect response. Select ten pairs of digits to represent the ten questions, allowing repeats. Record the number of pairs that are 01–25.

d. The estimated probability of getting more than half the questions correct by guessing is about 0.08.

E19. a. Each baby has probability 0.49 of being a girl. The sex of each baby is independent of that of the other babies in that family.

b. Assign pairs of digits 01–49 to a girl and pairs 50–99 and 00 to a boy. Select pairs of digits, allowing repeats, until a pair in the range 01–49 is selected. Record the number of pairs needed.

d. The estimated average number of babies per family for families who keep having babies until they have a girl is about 2.1.

E21. a. Each backpack is equally likely to be picked up by each friend. The backpacks are selected independently of one another.

b. Let backpack 1 go with person 1, backpack 2 with person 2, and so on. Assign digits 1 and 2 to backpack 1, 3 and 4 to backpack 2, 5 and 6 to backpack 3, 7 and 8 to backpack 4, 9 and 0 to backpack 5. Select a digit to represent the backpack taken by person 1. Select digits for the other four people, skipping digits that represent a backpack already picked. Record whether each person gets his or her own backpack.

d. The probability that each person gets his or her own backpack is 0.008.

E23. a. You grab a key at random. Each key is equally likely to be picked on any one grab.

b. Assign the digit 1 to your car key and the digits 2–6 to not your car key. Select two
digits, skipping 7, 8, 9, 0, and any repeats, and record whether the second digit is 1.

d. The probability that if you grab one key at a time you will get your key on the second draw is \( \frac{1}{6} \).

E25. a. You are assuming that the prize is equally likely to be behind any door and that Monty knows which door it is.

b. Assign digits 1–3 to door 1, digits 4–6 to door 2, and digits 7–9 to door 3. Ignore 0. Select a digit at random to be the door hiding the car. Select a second digit at random to be the door you choose. If you choose the door with the prize, there are two doors Monty might open. In this case, select a third digit at random to determine which door Monty opens. Let digits 1–5 mean that he opens the door to the left and digits 6–9 and 0 mean that he opens the door to the right. Record whether switching to the other unopened door results in the car or a goat.

d. The probability of winning the car with the switching strategy is \( \frac{2}{3} \).

Section 5.3

P14. a. no; yes  
b. 0.266  
c. no

d. 3,419,000

P15. b. about 0.3825

P16. a. no  
b. yes  
c. yes  
d. no

P17. a. yes  
b. 0.60

c. \( \frac{5}{16} \)

d. \( \frac{6}{16} \)

d. These events are not disjoint.

P19. a. no

b. \( \frac{24,291 + 19,755 + 67,331}{231,459} \approx 0.4812 \)

c. \( \frac{44,046 + 87,086 - 19,755}{231,459} \approx 0.4812 \)

P20. a. \( \frac{6}{36} + \frac{3}{36} - \frac{1}{36} = \frac{8}{36} \)

b. \( \frac{6}{36} + \frac{6}{36} - 0 = \frac{12}{36} \)

c. \( \frac{6}{36} + \frac{1}{36} = \frac{11}{36} \)

P21. \( \frac{1}{2} + 1 - 1 = \frac{3}{4} \)

P22. \( \frac{30}{36} + \frac{5}{36} - \frac{4}{36} = \frac{31}{36} \)

E27. The events in a and c are disjoint.

E29. a. 0.511; 0.0085; 0.515

b. 0.489; 0.210; 0.692  
c. 0.017; 0.667

d. \( \frac{190}{216} \) or about 0.88

e. \( \frac{20}{216} \)

Section 5.4

P23. a. \( \frac{470}{2201} \)  
b. \( \frac{344}{711} \)  
c. \( \frac{1731}{2201} \)  
d. \( \frac{367}{711} \)  
e. \( \frac{367}{1731} \)

P24. a. 0.027  
b. 0.028

c. A white hourly worker is a bit more likely to be paid at or below minimum wage than is an hourly worker in general.

d. 0.131  
e. 0.117

f. A worker who is paid at or below minimum wage is a bit less likely to be black than is an hourly worker in general.

P25. a. \( \frac{1}{2} \)

b. \( \frac{3}{4} \)

c. \( \frac{1}{3} \)

d. \( \frac{1}{3} \)

P26. a. \( \frac{1}{2} \)

b. \( \frac{1}{13} \)

c. \( \frac{1}{4} \)

P27. b. \( \frac{711}{2201} \)  

P31. P\text{doubles} = \frac{1}{5}  
P\text{sum of 8} = \frac{1}{6}

P33. 56.7%

P34. b. About 69% of samples that the technician decides are contaminated actually are contaminated.

c. About 97% of samples that the technician decides aren't contaminated actually aren't contaminated.

d. PPV = 0.90; NPV = 0.90

P35. PPV = 0.75; NPV = 0.20; sensitivity = 0.60; specificity \( \sim 0.33 \); no
P36. a. \( \frac{1}{2} \) 
   b. 1 
   c. The probability depends on the probability model used for selecting the coin.
   d. same as part c 
   e. 1 
   f. 0 

E41. a. 0.2883  
   b. 0.2420  
   c. 0.1367 
   d. same as part c 
   e. 1 
   f. 0 

E43. a. 0.3548 
   b. 0.1367 

E45. a. \( \frac{3}{5} \) 
   b. \( \frac{2}{5} \) 
   c. \( \frac{1}{2} \) 
   d. \( \frac{1}{2} \) 
   e. \( \frac{3}{4} \) 
   f. \( \frac{1}{4} \) 

E47. a. 0.0059 
   b. 0.1367 
   c. 0.0059 
   d. 0.0060 
   e. \( \frac{11}{52} \), or \( \frac{1}{4} \) 
   f. \( \frac{12}{51} \) 

E49. a. 24% 
   b. This is a new test, so it is unlikely that false negatives have shown up yet. 
   c. 177 

E51. b. 0.088  
   c. 0.661 
   d. no 
   e. no; no 
   f. yes 

E53. a. no 
   b. \( \frac{1}{3} \) 
   c. \( \frac{1}{2} \) 
   d. \( \frac{2}{3} \) 
   e. 1 
   f. 0 

Section 5.5

P37. \( \frac{1490}{2201} \neq \frac{1364}{1731} \), so not independent; no 

P38. a and c 

P39. b. 0.4872 

P40. a. 0.9957  
   b. 0.9998 
   c. 0.0121 
   d. 0.1958 

E57. only b 

E59. a. \( \frac{12}{100} \), \( \frac{37}{100} \), \( \frac{6}{12} \), \( \frac{6}{36} \) 
   b. no 
   c. No; six people are both. 

E61. a. BB, BG, GB, and GG 
   b. two boys; two girls 
   c. 0.2601; births are independent. 

E63. a. 0.021  
   b. 0.449 
   c. 0.753 

E67. a. 0.11^{12} \approx 3.14 \times 10^{-12}  
   b. 0.753 

E69. It is not reasonable to assume that the results of two surgeries on the same person at the same time are independent, so the probability is unknown. 

E71. 21.4% 

E75. b. \( P(win) = P(win \mid day) \cdot P(day) + P(win \mid not day) \cdot P(not day) = \frac{11}{21} \cdot \frac{21}{78} + \frac{30}{57} \cdot \frac{57}{78} = \frac{41}{78} \) 

d. \( P(win \mid day) = \frac{P(day \mid win) \cdot P(win)}{P(day \mid win) \cdot P(win) + P(day \mid not win) \cdot P(not win)} = \frac{\frac{11}{21} \cdot \frac{41}{78} + \frac{41}{78} \cdot \frac{57}{78}}{\frac{11}{21} \cdot \frac{41}{78} + \frac{30}{57} \cdot \frac{57}{78}} = \frac{11}{21} \) 

Chapter Summary

E77. a. 24  
   c. \( \frac{4}{24} \)  
   d. \( \frac{2}{24} \) 
   e. no; no  
   f. no; yes 

E79. a. Randomly select seven digits from 1 to 9 (with replacement). See if the number 3 is among the seven digits selected. If so, the run was a “success.” Repeat many times. 
   b. 0.562 

E81. a. 0.675  
   b. 0.025  
   c. 0.975  
   d. 0.75 

E83. a. \( \frac{1}{6} \), or about 0.167 
   b. 0.249999975 
   c. no; no; part b, because of the large population size compared to the sample size 

E85. a. 0.381; 0.310; 0.286; 0.405; 0.923; 0.75 
   b. no; \( P(Republican) \neq P(Republican \mid voted \ Republican) \) 
   c. no 

E87. b. no; \( P(B) = \frac{5}{16} \neq P(B \mid A) = \frac{1}{2} \) 

E89. This is a reasonably good test, but if depression is a serious impediment to recovery, a test should have higher sensitivity than 86%. The number of false positives, 8, isn’t much of a problem because these people will receive further testing. 

E91. a. about 3000  
   b. about 0.3 

Chapter 6

Section 6.1

P1. 2, 4 

P2. Let any digit represent smoking is not responsible. The other nine digits represent smoking is responsible. 

P3. For example, begin by assigning the triples 001–524 to represent families with zero children. Then the first triple, 488, represents a family with no children. 

P4. 0.007744, if it’s reasonable to assume independence; very close 

P5. \( P(X = 1) = \frac{11}{36}, P(X = 2) = \frac{9}{36}, P(X = 3) = \frac{7}{36}, P(X = 4) = \frac{5}{36}, P(X = 5) = \frac{3}{36}, P(X = 6) = \frac{1}{36} \)
P6. $P(X = 0) = 0.027; P(X = 1) = 0.189;
     P(X = 2) = 0.441; P(X = 3) = 0.343$

P7. b. $\frac{1}{10}$
    c. $P(X = 0) = \frac{1}{10}; P(X = 1) = \frac{6}{10}; P(X = 2) = \frac{3}{10};
     P(X = 3) = 0$

P8. 0.873; 1.096

P9. sum: triangular and symmetric, 7, 2.415; larger number: skewed left, 4.472, 1.404

P10. a. $P(X = 0) = \frac{3}{15}; P(X = 1) = \frac{9}{15}; P(X = 2) = \frac{3}{15}$
     b. 1; 0.632
     c. $\frac{12}{15}$

P11. a. $1.20$ b. $3.60$

P12. $116$

P13. a. $33$
     b. $19.52$ (after including the probability 0.2 of no rental); yes

P14. a. triangular
     b. 0:5.833
     c. $3.5 - 3.5 = 0; 2.917 + 2.917 = 5.834$
     (The difference is due to rounding.)

P15. $1060.80; 106.58$, assuming savings from week to week are independent

P16. a. 349
     b. It would not be unusual to be off by 14.057 vehicles or so.
     c. That the numbers of vehicles in the households are independent of each other;
     however, people living in the same apartment complex probably will have similar income
     levels and so are more likely to have closer to the same number of cars.

P17. 3.49; 1.41; quite close

P18. a. 2.236; 1.115 b. 4.472; 1.577

E1. a. 2.16; 2.221
     b. Assign random digits by pairs.

E3. 126.5; 29

E5. $P(X = 0) = 0.729; P(X = 1) = 0.243; P(X = 2) = 0.027; P(X = 3) = 0.001$

E7. $P(X = 0) = 0.6724; P(X = 1) = 0.2952; P(X = 2) = 0.0324$

E9. a. $209.50; 4790.50$ b. $214.50$
     c. for example, whether there is a burglar alarm

E11. a. for example, $P(X = 0) = 0.0077$,
     $P(X = 5) = 0.1440$
     b. quite close
     c. $3.49$ is equal to the mean found in P17 and is close to the mean of the simulated
distribution, 3.53.

E13. a. $0.185$ b. $315,000$ c. 37%

E15. a. Brand A: $1115; Brand B: $1200
     b. The advantage of buying Brand A is that you expect to pay less and there is a 70% chance you will pay less with Brand A than you will with Brand B, even with one repair. The advantage of buying Brand B is that the maximum you might spend is less.

E17. a. $\frac{2}{16}$ b. $\frac{2}{16}$ c. 1.25

E19. a. $P(X = 2) = 0.02; P(X = 3) = 0.14; P(X = 4) = 0.36; P(X = 5) = 0.48$
     b. 4.3 months; 0.781 month
     c. 2.5 + 1.8 = 4.3

E21. a. for example, $P(X = -2) = \frac{2}{24}; P(X = 3) = \frac{3}{24}$
     b. $\mu_1 - \mu_2 = 3.5 - 2.5 = 1; \mu_3; \sigma_1^2 + \sigma_2^2 = 2.917 + 1.25 = 4.167 = \sigma_3^2$

Section 6.2

P19. for example, $P(X = 3) = 0.3125$ and
     $P(X = 6) = 0.015625$

P20. 0.2188; 0.1094; 0.0352

P21. 0.1641; 0.5

P22. a. 0.6309 b. 0.3691
c. for example, $P(X = 3) = 0.005668$ and
     $P(X = 5) = 0.000005$

P23. For example, assign the triples 001–088 to represent the dropouts and the other triples to represent non-dropouts.

P24. $P(X = 0) = 0.0053; P(X = 1) = 0.0575; P(X = 2) = 0.2331; P(X = 3) = 0.4201;$
     $P(X = 4) = 0.2840$

P25. a. 0.1875 b. 8; 2; 0.227
c. that the sample was not selected randomly

P26. Five-question quiz; the probability of getting 3 or more correct on a 5-question quiz is 0.5. On a 20-question quiz, the probability of getting 12 or more correct is only about 0.25.

E23. a. 0.4019 b. 0.0322
c. 0.5981 d. 0.8038

E25. a. 0.8446 b. 3.175; 1.6649

E27. a. −$6.00$ b. 0.0571 c. $9.20$
E29. a. 0.973  b. 0.99927; no  c. 4
E31. a. 7  b. 17
E33. If 100 different polls each randomly selected three adults, you would expect to see a total of 81 college graduates, or an average of 0.81 college graduate per poll.

Section 6.3

P27. a. $\frac{1}{6}$  b. $\frac{5}{36}$

P28. a. 0.0191  b. 0.0225

P29. a. 0.0074  b. 0.9993

P30. a. 0.3  b. 0.21  c. 0.1029  d. Whether you get through is independent of the outcome on previous calls, and each call still has probability 0.3 of being completed.

P31. a. 10  b. 9.4868  c. 20; 30

P32. a. 1.1765  b. 0.4556  c. 11.765  d. 1.441

E35. a. $\frac{1}{4}$  b. $\frac{1}{64}$  c. $\frac{4}{3}$  d. 0.5

E37. a. 1.1364; 3.4091  b. 0.3936  c. $\$113.64$; $\$39.36$  d. no; only 1.44% of the time

E39. 0.5528; this assumes the two test results are independent, which is almost certainly not the case.

E41. a. The probabilities for a geometric distribution follow this same pattern: $p, qp, q^2p, q^3p, q^4p, \ldots$  
b. $\frac{p}{1-q} = \frac{p}{p} = 1$

Chapter Summary

E43. a. $\frac{3}{144}$  b. 6.5; 3.452  c. 13; 4.882

E45. a. $P(X = 1) = 0.19; P(X = 2) = 0.45; P(X = 3) = 0.36$

b. $P(X = 1) = 0.01; P(X = 2) = 0.15; P(X = 3) = 0.84$

c. 2.17 months; 2.83 months; System II

d. 0.6513; the probability that a donation is type B is independent of the type of the donations previously checked.

b. 0.5487  c. No; the probability is only 0.0399. Check more donations.

d. 0.729

E49. a. 0.9974; that in your region 80% of the girls want to do their best in all classes  
b. 0.8782; that in your region 65% of the boys want to do their best in all classes  
c. 0.9979; that the sample is about half girls

E51. a. $\$31.804$  b. about 128

E53. a. 300, 450, 900, 325, 475, 925, 700, 850, and 1300; 313.8028; 98,472.22  
b. 33,472.22; 65,000; 33,472.22 + 65,000 = 98,472.22  
c. no; 182.9542 + 254.9510 = 313.8028  
d. yes; 341.67 + 350 = 691.67  
e. no; 225 + 250 = 700

E55. a. 2  b. 0.9990  c. 1.001 flips; less, because there is a smaller chance of all ten living than there is of a single flipper living  
d. about 4.8 rounds

Chapter 7

Section 7.1

P1. a. 50; 7300  
b. Randomly select three planets. Sum the numbers of moons. Repeat many times.

P2. a. I—B; II—A; III—C  
b. no, because a mean of 86 or greater did not occur in 100 random samples in Histogram C

P3. a. D; B; C  b. D; A

P4. a. $\$8 million$; $\$1.9 million$  
b. There are 15 possible amounts, each with probability $\frac{1}{15}$.

c. $\frac{12}{15}$

P5. a. $\mu = 261.8; \sigma = 171.85$  
b. Both equal 261.8.

c. The $SE$, 105.23, is much less than $\sigma = 171.85$.

d. No; the mean of the sampling distribution of the range is much less than the actual range, 471.

e. yes
E1. 
   a. A—20; B—1; C—2; D—50
   b. The means are all about 4.5, regardless of sample size.
   c. The larger the sample size, the smaller the spread.

E3. 
   b. compared to mean: more skewed left, centered higher, about equal spread
   c. good estimator of the midrange of the population, but tends to be too large compared to the mean and median

E5. 
   b. 62, 63; 62, 64; 62, 65; 63, 64; 63, 64; 63, 65; 64, 64; 64, 65; 64, 65
   c. Sampling distribution: mean 63.6, SE 0.6245; population: mean 63.6, SD 1.02; means are equal, but SE < SD.
   d. 63, 64, 64, 65, 64, 65, 64, 65, 64, 65
   e. 1, 2, 2, 3, 1, 1, 2, 0, 1, and 1

E7. 
   a. Unbiased; the mean of the sampling distribution of the sample mean is equal to the population mean.
   b. biased; tends to be too small
   c. biased; tends to be too small

E9. 
   a. The sample of 48 packages will contain 0, 24, or 48 lb of spoiled fish, but the probabilities are unknown.
   b. No, because the sample size was only 2; sample one fish (or one package) from each of a larger number of cartons.

E11. 
   e. 4; 1\frac{1}{3}; 2\frac{2}{3}; f. \frac{8}{3}
   g. by \( n - 1 \), because the average of the sample variances \( \frac{8}{3} \) is equal to the population variance
   h. too small

E13. 
   a. \( \frac{1}{3}N \)
   b. at \( \frac{4}{3}N \)
   c. \( N = \frac{5}{4} \) (sample maximum)

Section 7.2

P7. 
   a. I—population; II—\( n = 10 \); III—\( n = 4 \); mean 1.7 for each; SD for I, II, and III about 1, 0.3, and 0.5, respectively
   b. 1.7; yes
   c. 0.5; 0.316; yes
   d. The population is slightly skewed, and the two sampling distributions get more nearly normal as the sample size increases; yes.

P8. approximately normal; .66; .009553

P9. 
   a. approximately normal; mean 0.9; SE 0.035
   b. from 0.83 to 0.96

P10. 
   a. 0.3058 (0.3050 using Table A)
   b. 0.1549 (0.1539 using Table A)
   c. 0.15
d. −2.4 and 15.4

P11. 
   a. 0.04
   b. 0.0384
   c. 0.64
d. 0.6384
   e. 0.36
   f. 0.3616

P12. would not change, because the sample size in both cases is a very small fraction of the population size

P13. 
   a. 0.1226 (0.1230 using Table A)
   b. 0.1635 (0.1645 using Table A)

P14. 
   a. No; the probability is only 0.002.
   b. Yes; the probability is 0.982.

P15. 
   a. (0.469, 1.331)
   b. (0.684, 1.116)
   c. (0.832, 0.968)
   d. (0.866, 0.934)

E15. 
   a. I: B, 25; II: A, 4; III: C, 2
   b. yes
   c. The distributions for \( n = 4 \) and \( n = 2 \) reflect the skewness of the population. The distribution for \( n = 25 \) looks approximately normal.
   d. works well for \( n = 25 \) and less well for \( n = 4 \) and \( n = 2 \)

E19. 
   a. No; two or more accidents happened in about 27% of the days.
   b. The top plot is for 8 days.
   c. No; an average of 1.75 or more accidents occurred about 16 times out of 200.
   d. Yes; an average of 1.75 or more accidents occurred about 4 times out of 200.
   e. It isn’t reasonable to assume that the days are independent; for example, if the first of the days has ice or snow, it’s likely the other days will also.

E21. 
   a. 0.4602
   b. 0.4207
   c. 0.3085
   d. Assign the pairs of digits 01 through 46 to be a score of 510 or greater. The other pairs represent a score lower than that. Take four pairs of random digits and see if all four represent scores of 510 or greater.

E23. 
   a. 0.4729 (0.4714 using Table A)
   b. about 384
E25. a. 134.8 b. almost 0 c. The small number can’t reasonably be attributed to chance.

E27. a. 27 b. $4025

E29. a. remains at the population mean, \( \mu \), for all sample sizes; shrinks by a factor of \( \frac{1}{\sqrt{n}} \)
b. increases by a factor of \( n \); stretches by a factor of \( \frac{1}{\sqrt{n}} \)

E31. a. 1021; 161.23 b. 0.0852 (0.0853 using Table A)
c. mean 1021; SE unknown, because the two scores aren’t independent

d. 0.2378 (0.2389 using Table A)
ed. (705, 1337)
e. (0.448, 0.752) b. (0.504, 0.696)
c. (0.552, 0.648)

Section 7.3

P16. 0.853

P17. a. \( n = 100; n = 10 \); yes b. 0.10 in each case
c. \( n = 10; n = 100 \) d. 10

P18. a. means 0.10; SEs 0.095, 0.067, 0.047, 0.03
b. 0.0999999; 0.095; yes
P19. a. approximately normal; 0.53; 0.05 b. no, because the probability of getting nine or fewer women just by chance is close to 0

P20. a. close to 0 b. yes
P21. a. (0.448, 0.752) b. (0.504, 0.696)
c. (0.552, 0.648)

E35. a. approximately normal, mean 0.92, SE 0.00858 b. approximately normal, mean 920, SE 8.58
c. 0.0099
d. 0.2800 (0.2810 using Table A)
e. 937 or more or 903 or fewer; larger than 0.937 or smaller than 0.903

E37. a. slightly skewed left; mean still 0.92; SE larger b. slightly skewed left; mean one-tenth as large; SE smaller
c. larger

E39. a. 0.5 b. 0.000011 (close to 0 using Table A)
c. No; this group is special in that the employees are all old enough to have a job.

E41. 0.1191 (0.1188 using Table A); you are assuming that the 75 married women were selected randomly and independently from a large population in which 60% are employed.

E43. a. yes; no b. For a fixed \( p \), the SE decreases as \( n \) increases. For a fixed \( n \), the closer \( p \) is to 0.5, the larger the SE.
c. For a fixed \( p \), the skewness decreases as \( n \) increases. For a fixed \( n \), the farther \( p \) is from 0.5, the more skewness.
d. for \( n = 25 \) or 100 when \( p = 0.4 \), for \( n = 100 \) when \( p = 0.2 \)

E45. a. \( \bar{x} = 0.65 \), which is equal to the proportion of successes, \( \hat{p} = \frac{26}{40} \)
b. \( \mu = 0.6; \mu_{\hat{p}} = 0.6 \), which is the same
c. The sample proportion is a type of sample mean, if successes are assigned the value 1 and failures the value 0.

Chapter Summary

E47. a. Assign each city a two-digit number from 01 to 32. Generate pairs of random digits until you get three distinct pairs between 01 and 32. Ignore 33–99 and 00 and repeats.
b. 140.9; 68.94 c. no

E49. a. approximately normal; 0.68; 0.033 b. 0.17
c. less than 0.615 or greater than 0.745 d. 0.8185 (0.8186 using Table A)

E51. a. regular increase from 40 to 60 followed by a sharp drop after 60 and another sharp drop after 65 b. 0.6048 (0.6065 using Table A)

E53. a. approximately normal, mean 4, standard error 3.680 b. 0.7066 (0.7054 using Table A)
c. 0.1385 (0.1379 using Table A)

E55. a. Select two integers at random from 0 through 9 and compute their mean. Repeat. Subtract the second mean from the first. You win that amount.
b. shape slightly rounded-out triangular, mean 0, SE 2.03
E57. a. becomes more approximately normal; stays at \( p \); decreases
b. becomes more approximately normal; increases; increases
E59. 0.739
E61. a. 12.8 and 4.56; 8 and 4
b. \( \mu_{\text{total}} = 20.8 \) and \( \sigma_{\text{total}}^2 = 6.96 \)
c. 12.8 + 8 equals 20.8, as you would expect; but 4.56 + 4 is not equal to 6.96, and you should not expect it to be because the morning and afternoon times were not selected independently when you computed 6.96.

Chapter 8
Lesson 8.1
P1. a. 0.264 and 0.536 b. yes
P2. Let random digits 1, 2, and 3 represent successes. Count the number of successes in 40 digits.
P3. about 22–42
P4. about 14–26
P5. no
P6. about 28–39; about 0.70–0.975
P7. about 15% to 40%
P8. about 30% to 60%
P9. a. about 50% to 75% b. (0.475, 0.775)
P10. a. Conditions are met.
   b. (0.612, 0.688); 0.038
   c. (0.618, 0.682); 0.032
   d. 95%, because a longer confidence interval has a greater chance of including \( p \)
P11. a. Conditions are met. b. (0.024, 0.056)
   c. narrower, because \( \hat{p} \) is farther from 0.5
P12. 72
P13. 1
P14. The horizontal line segments will be shorter; the confidence intervals will be narrower.
P15. 1.28; smaller, because \( z^* \) is smaller
P16. 9 times as large
P17. a. 2,401; 16,590; 27,061
P18. 879
P19. a. The second two conditions are met, but random sampling isn't mentioned.
b. (73.3%, 78.7%)
c. the proportion of all students ages 12–17 with Internet access who would say this
d. If you could ask all students ages 12–17 with Internet access whether they go online to get news or information about current events, you are 95% confident that the proportion who would say yes would be somewhere in the interval from 73.3% to 78.7%.
e. Suppose you could take 100 random samples from this population and construct the 100 resulting confidence intervals. You'd expect the proportion of all students ages 12–17 with Internet access who would say this to be in 95 of these intervals.
E1. a. no b. about 0.65; about 0.35
c. about 30% to 55%
E3. a. about 0.05–0.25 b. (0.039, 0.261)
c. They are similar. The chart would give a more exact interval but doesn't include all values of \( p \). Also, \( n(1 - \hat{p}) < 10 \), so the normal approximation used in part b is not reliable.
E5. a. no apparent randomization
   b. no, because there is no indication of a random sample
   c. yes
   d. the proportion of all U.S. teens who know this
e. Suppose you could take 100 random samples from this population and construct the 100 resulting confidence intervals. You'd expect the proportion of all U.S. teens who know this to be in 95 of these intervals.
E7. Conditions have been met. (0.485, 0.555)
E9. No; Display 8.2 applies only to samples of size 40.
E11. Conditions are met. You are 95% confident that, if you asked all adults, the percentage who thought this would be between 78.6% and 83.4%.
E13. a. gets narrower b. gets wider
E15. a. yes, because 0.042 rounds up to 5%
   b. 1068
E17. A, D, F, H
E19. a. about 0.013 using 95% confidence
   b. at least 212,345
E21. a. You want to maximize \( p(1 - p) \); because \( p \) is restricted to \([0, 1]\).
   b. quadratic, or parabola opening down
   c. \( x = \frac{1}{2} \) and \( y = \frac{1}{4} \)

Section 8.2

P20. a. \( p = \frac{2}{3} \), where \( p \) is the proportion of today’s teens who want to study more about medical research
   b. 0.575
   c. not statistically significant
P21. a. \( p = 0.5 \), where \( p \) is the probability that the student gets an answer correct.
   b. 0.75
   c. statistically significant
d. No; the student might have been really lucky.
P22. a. \( p = 0.75 \), where \( p \) is the proportion of all juniors who will want extra tickets
   b. 0.80
c. not statistically significant
d. no
P23. −1.23
P24. 0.149
P25. If spinning a penny is fair, there is only a 0.0016 chance of getting 10 or fewer heads or 30 or more heads in 40 spins.
P26. If the cats could not distinguish between the rakes, the probability of getting the rake with the food 28 or more times (or 22 or fewer times) is 0.3961.
P27. a. \( H_0: p = 0.69 \), where \( p \) is the proportion of houses in your community that are occupied by their owners; \( H_1: p \neq 0.69 \)
   b. −1.376
   c. If the proportion of houses in your community that are occupied by their owners is 69%, then the probability of getting 30 or fewer or 39 or more owner-occupied houses in a random sample of 50 houses is about 0.1688.
d. With a \( P \)-value this high, there isn’t sufficient evidence to reject \( H_0 \), that the proportion of houses occupied by their owners in your community is 0.69.
P28. B

P29. a. \( z = -5.01 \) is more extreme than \( z^* = \pm 1.96 \); reject \( H_0 \), that spinning a quarter is fair.
b. \( z = 1.34 \) is less extreme than \( z^* = \pm 1.96 \); don’t reject \( H_0 \), that flipping a quarter is fair.
P30. a. about \( \pm 1.55 \) b. 0.0836
P31. a. 1% b. \( \pm 2.33 \)
P32. Conditions are met. \( H_0: \) The probability of heads when a quarter is spun is 0.5. \( H_1: \) The probability is not 0.5. Because \( z = -5.01 \) is more extreme than \( z^* = \pm 1.96 \), reject \( H_0 \) that spinning a quarter results in heads half the time.
P33. Conditions are met. \( H_0: \) The proportion of all U.S. bookstores that sell DVDs is 0.5. \( H_1: \) The proportion is not 0.5. Because \( z = 1.34 \) is less extreme than \( z^* = \pm 1.96 \), the result from this sample is reasonably likely to occur if half of all bookstores sell DVDs. Do not reject \( H_0 \).
P34. a. \( p = 0.2 \), where \( p \) is the probability that the student gets any one answer correct
   b. \( p = 0.5 \), where \( p \) is the proportion of newscasters who are men (or women)
c. \( p = \frac{1}{7} \), where \( p \) is the proportion of weekly car-washing people who wash on Saturday
P35. B
P36. a. With \( z = 1.43 \), Hila should not reject \( H_0 \).
b. no
P37. a. Yes, they each have a 0.05 chance.
b. no, because \( H_0 \) isn’t false
P38. a. no, because \( H_0 \) isn’t true
   b. Both could, but Jeffrey is more likely to because he has the smaller sample size.
P39. \( \alpha = 0.10 \)
P40. \( n = 200 \)
P41. 10%, which is farther from \( p_0 \)
P42. a. Random sample; \( np_0 = 500 \) and \( n(1 - p_0) = 500 \) are both at least 10; the number of U.S. adults is at least 10,000.
b. \( H_0: \) The proportion \( p \) of all U.S. adults who would say they are satisfied is 0.5. \( H_1: p < 0.5 \).
c. −2.53; 0.0057
d. The \( P \)-value, 0.0057, is less than \( \alpha = 0.05 \), so reject \( H_0 \). There is sufficient evidence to support the claim that the proportion of all U.S. adults who would say they are satisfied is less than 50%.
P43.  

a. Random sample; \( np_0 = 500 \) and 
\( n(1 - p_0) = 500 \) are both at least 10, and 
the number of U.S. adults is at least 10,000.

b. \( H_0: \) The proportion, \( p \), of all U.S. adults 
who would say they are dissatisfied is 0.5. 
\( H_1: p > 0.5 \).

c. 0.632; 0.2635

d. The \( P \)-value, 0.2635, is more than \( \alpha = 0.05 \), 
so do not reject \( H_0 \). If the proportion of U.S. 
adults who would say that they are dissatisfied 
is 0.5, you are reasonably likely to get a 
random sample of 1000 adults in which 
51% say they are dissatisfied.

P44.  
Conditions are met if the number of teens in your 
community is at least 1690. \( H_0: \) 51% (or even 
less) of teens in your community know this.

\( H_1: \) The percentage is greater than 51%. Because 
the \( P \)-value, about 0.1473, is greater than 0.05, 
don't reject \( H_0 \). If it is true that the percentage 
of teens in your community who know this is 51%, 
then you are reasonably likely to get 55% who 
know this in a random sample of 169 teens.

P45.  

a. false  

b. false  

c. false  

d. true

E25.  
D

E27.  

a. \( H_0: p = 0.6 \), where \( p \) is the proportion of all 
students in the school who carry backpacks 
to class

b. \( z = -0.6455 \); if it is true that 60% of the 
students carry a backpack to class, the 
probability of getting 26 or more or 22 or fewer 
in a random sample of size 40 is about 0.519.

c. Because the \( P \)-value is more than 0.05, do 
not reject \( H_0 \). The results from the sample are 
consistent with an overall percentage of 60% of 
students carrying a backpack to class.

E29.  
Conditions are met. \( H_0: p = 0.25 \), where \( p \) is 
the proportion of all U.S. adults who would pick 
a cure for cancer. \( H_1: p \neq 0.25 \). Because the 
\( P \)-value, about 0.0002, is less than 0.05, reject \( H_0 \).
In a random sample from a population with 25% 
successes, there is almost no chance of getting 
as many successes as Gallup got in this sample. 
Thus, it is not plausible that 25% of all adults 
would pick a cure for cancer.

E31.  
Conditions are met. \( H_0: p = \frac{2}{3} \), where \( p \) is 
the proportion of all U.S. adults who would have 
said this. \( H_1: p \neq \frac{2}{3} \). The \( P \)-value, 0.08, is larger 
than 0.05, so don't reject \( H_0 \). You have insufficient 
evidence to conclude that if all U.S. adults had 
been asked, the proportion who would have said 
this would be different from \( \frac{2}{3} \).

E33.  

5

E35.  

a. yes for both; Taline  

b. no, because \( H_0 \) isn't false

E37.  

a. For example, the probability of exactly 4 heads 
is 0.00462.

b. 5 or fewer, or 15 or more

c. 0.1272  

d. Increase the sample size.

E39.  

a. no  

b. Yes, 0.50 is a plausible value of \( p \).

c. (0.365, 0.657), which contains 0.5

d. A smaller margin of error means greater 
power.

e. Use a larger sample.

E41.  

a. one-sided  

b. a random sample  

c. almost 0

b. One of three conclusions could be drawn: This 
area always could have had a larger proportion 
of birds with the mutation, there might have 
been something about the mutation that made 
those birds more likely to be in the sample, 
or the proportion of birds with the mutation 
is higher since the radiation leak. If scientists 
can rule out the first two conclusions, they can 
go with the third.

e. Not really; a much smaller sample would 
have been sufficient because 66.8% is so far 
from 50%.

E45.  

a. one-sided

b. \( \hat{p} = 0.668; p_0 = 0.5 \)

c. yes, except for rounding error

d. yes

e. Larger sample sizes make it easier to reject a 
false \( H_0 \).
Section 8.3

P46. a. Conditions are met.
   b. You are 95% confident that the difference in the two rates, 1994 − this year, is between −0.206 and 0.066.
   c. Yes. There is insufficient evidence to support a claim that there was a change in the percentage.

P47. a. Conditions are met, because Gallup uses a randomization that can be approximated by random sampling.
   b. 0.14 to 0.30
   c. You are 95% confident that the difference between the proportion of all 13- to 15-year-olds who respond yes and the proportion of all 16- to 17-year-olds who respond yes is between 0.14 and 0.30.
   d. No. It is not plausible that there is no difference between these proportions.

E47. B and C

E49. a. Conditions are met except whether this is a random sample.
   b. (0.009, 0.571)
   c. You are 95% confident that if the judges had been given a choice of two dogs for each San Diego dog owner, the difference in the proportion of correct guesses for all purebreed owners and the proportion of correct guesses for all mutt owners would be between 0.009 and 0.571.
   d. No. This implies that the difference in the proportions probably was not due to chance, but it is a close call.
   e. Take bigger samples to be sure 0 isn’t in the confidence interval.

E51. a. Conditions are met.
   b. (−0.19, −0.03)
   c. You are 99% confident that the difference in the percentage of all men and the percentage of all women who prefer to be addressed by their last name is between −0.19 and −0.03.
   d. No. You are convinced that there is a difference between the percentage of women and the percentage of men who prefer to be addressed by their last name.

E53. Conditions are met. Because 0 isn’t included in the confidence interval (0.094, 0.126), it is acceptable to use the term “significantly more likely.”

E55. Generally, the length of the interval gets smaller.

E57. The sampling distribution of \( \hat{p}_1 - \hat{p}_2 \) is approximately normal.

E59. No, because the respondents weren’t selected independently from two different populations; use a confidence interval for a single proportion, which is (0.140, 0.304).

Section 8.4

P48. a. 0  b. 0.0154  d. 0.0006

P49. a. false  b. true  c. true  d. true  e. true

P50. Conditions are met if the town has at least 7500 probable voters. \( H_0: \) The proportion, \( p_1 \), of all potential voters who favored the candidate at the time of the first survey is equal to the proportion, \( p_2 \), of all potential voters who favored the candidate one week before the election.
   \( H_a: p_1 > p_2. \) Do not reject \( H_0. \) If there is no difference between the proportion of potential voters who favored the candidate at three weeks and the proportion at one week, then there is a 0.1741 chance of getting a difference of 0.0257 or larger with samples of these sizes.

P51. Type I

E61. a. 0.04  b. 0.0585  d. 0.432 (0.4325 using Table A)

E63. a. It’s reasonable to assume that a larger tumor is more likely to spread than is a smaller tumor.
   b. no, because these are not independent random samples of patients with tumors of different sizes
   c. yes
   d. If large tumors and small tumors are equally likely to metastasize, then there is about a 0.02 probability of seeing a difference in proportions at least as large as that seen in this study.

E65. a. Conditions are met. \( H_0: \) The proportion, \( p_1 \), of all girls who would have said yes is equal to the proportion, \( p_2 \), of all boys who would have said yes. \( H_a: p_1 > p_2. \) Reject \( H_0. \) If the proportions of all girls and all boys ages 12–17 who would have said yes were equal, the probability of getting a difference in proportions of 13% or greater from samples of these sizes would be only 0.0007.
b. Not all boys and girls are equally likely to be in the sample.

c. That about half of teens ages 12–17 are girls and half boys; you are 95% confident that if you asked all teens this question, between 65.5% and 73.4% would say yes.

E67. Conditions are met. $H_0$: The proportion, $p_1$, of all adult Americans who would give NASA a favorable rating in 2005 is equal to the proportion, $p_2$, who would give NASA a favorable rating in 1999. $H_1$: $p_1 > p_2$. With a $P$-value of 0.0008, reject $H_0$. There is strong evidence that NASA was being looked upon more favorably in 2005 than it was in 1999.

E69. a. Conditions are met. $H_0$: The proportion, $p_1$, of adults who logged on in 2003 is equal to the proportion, $p_2$, of adults who logged on in 2005. $H_1$: $p_1 < p_2$. The $P$-value, 0.037, is less than $\alpha = 0.05$, so reject $H_0$. You have sufficient evidence that the proportion has increased.

b. No; percentages less than 50% are in the 95% confidence interval (0.479, 0.541).

E71. B

E73. Conditions were checked. $H_0$: $p_1 - p_2 = 0.02$, where $p_1$ is the proportion of all males who are left-handed and $p_2$ is the proportion of all females who are left-handed. $H_1$: $p_1 - p_2 > 0.02$. The $P$-value, 0.284, is greater than $\alpha = 0.05$, so don’t reject $H_0$ that the difference in the two proportions is 2%.

Section 8.5

P52. Conditions are met. Suppose all the subjects could have been given aspirin and all could have been given the placebo. Then you are 95% confident that the difference in the proportion who would get ulcers is in the interval $0.0028 \pm 0.0031$. Because 0 is in this interval, it is plausible that there is no difference in the two proportions.

P53. a. yes  \hspace{1cm} b. $(-0.348, -0.062)$

c. Suppose it were possible that all the children could have been in both treatment groups. Then you are 90% confident that the difference in the proportion who would get arrested is in the interval $(-0.348, -0.062)$. Because 0 is not in this interval, it is not plausible that there is no difference in the proportions who would get arrested.

d. Although it seems like a big stretch to say that not going to preschool caused more children to get arrested, we have no other explanation other than an unlikely Type I error.

P54. a. Both conditions are met. $H_0$: If all patients could have been given Lipitor, the proportion, $p_1$, who had heart attacks would be equal to the proportion, $p_2$, who had heart attacks had they all been given Zocor. $H_1$: $p_1 \neq p_2$. Because the $P$-value, 0.082, is greater than $\alpha = 0.05$, don’t reject $H_0$. There is not sufficient evidence to conclude that, had all patients been treated with Lipitor, the proportion who had heart attacks would have been different than had all patients been treated with Zocor.

b. Both conditions are met. $H_0$: If all patients could have been given Lipitor, the proportion, $p_1$, who had heart attacks would be equal to the proportion, $p_2$, who had heart attacks had they all been given Zocor. $H_1$: $p_1 < p_2$. The test statistic is still $-1.74$. Because the $P$-value, 0.041, is less than $\alpha = 0.05$, reject $H_0$, that if all patients could have been given Lipitor the proportion who had heart attacks would be the same as the proportion who had heart attacks had they all been given Zocor. A one-sided test makes it easier to reject $H_0$ if the difference is in the direction $H_1$ states.

P55. Type II; Type I; a Type I error means the patient receives a more expensive drug even though it’s not more effective, and a Type II error means the patient is not given the more effective drug. Type II seems more serious.

P56. No randomization was done, so that condition is not met. $H_0$: The difference in the proportion of cold-turkey quitters who were successful and the proportion of planned quitters who were successful can be attributed to chance variation alone. $H_1$: The difference is too large to reasonably be attributed to chance. With a $P$-value so close to 0, reject $H_0$. Those who chose to quit cold turkey were more successful than the others, and this difference cannot reasonably be attributed to chance.

P57. a. observational study  \hspace{1cm} b. $(0.010, 0.058)$

c. Zero is not in this interval, so the difference in the proportion of people in this study
abused as children who later committed crimes and the proportion of people not abused as children who later committed crimes cannot reasonably be attributed to chance; no, because other contributing factors weren't controlled.

P58. greenways

E75. a. Conditions are met. You are 95% confident that if all the skiers in the experiment could have been in both treatment groups, then the difference in the proportions who got colds would have been in the interval (0.011, 0.187). Because 0 isn't in this confidence interval, you believe that vitamin C reduced the proportion who got colds.

b. Because 0 is in the interval (−0.016, 0.214), it is plausible that the same proportion of these skiers would have gotten a cold whether or not they took vitamin C.

E77. Conditions are met. You are 95% confident that if all the children had been given the placebo and if all had been given the vaccine, the difference in the proportions who developed polio would have been between −0.0005 and −0.0002. You believe that the vaccine reduced the proportion of children who developed polio. (Although this looks like a small difference, the vaccine cut the incidence of polio about in half.)

E79. Conditions are met. \( H_0: \) If all the men had been given aspirin, the proportion, \( p_1 \), who had a heart attack would have been equal to the proportion, \( p_2 \), who had a heart attack if all had been given the placebo. \( H_1: p_1 < p_2 \). With a \( P \)-value close to 0, reject \( H_0 \). If there is no difference in the proportion who would have had a heart attack had they all taken aspirin and the proportion who would have had a heart attack if they had all taken the placebo, then there is almost no chance of getting a difference of −0.0091 or smaller in the two proportions from a random assignment of treatments to these subjects.

E81. a. observational study

b. \( H_0: \) The difference between the proportion, \( p_1 \), of dementia-free people who exercise three or more times a week and the proportion, \( p_2 \), of those with signs of dementia who exercise three or more times a week can reasonably be attributed to chance variation. \( H_1: \) The difference cannot reasonably be attributed to chance. Because the \( P \)-value, 0.006, is less than \( \alpha = 0.05 \), reject \( H_0 \). No; there is evidence of an association between exercise and a delay of dementia for the group of persons in this study, but the reason isn't known.

E83. A difference as large as Reggie's between the regular season and the World Series would happen by chance to fewer than 17 players in 1000. Therefore, Reggie's record is indeed unusual and cannot reasonably be attributed to chance variation alone.

E85. a. 7; a value this small would occur only about 3.5% of the time if \( H_0 \) were true.

b. With a \( P \)-value of 0.035, reject \( H_0 \), which was also the conclusion in the example.

Chapter Summary

E89. a. Conditions are met. (0.754, 0.806)

b. You are 95% confident that if you were able to ask all teenagers with online access, between 75.4% and 80.6% would agree.

c. If you could repeat this survey with 100 different random samples of teens with online access, each of size 971, you would expect the (unknown) proportion of all teens with online access who agree with the statement to be in 95 of the resulting (probably different) confidence intervals.

E91. a. \( H_0: \) The difference in the proportion of snowboard injuries and ski injuries that were fractures can reasonably be attributed to chance variation. \( H_1: \) The difference is too large to be attributed to chance alone. With a \( P \)-value this close to 0, the difference in the proportions of injuries that were fractures is statistically significant.

b. No. You need to know the total number of snowboarders and skiers.

E93. a. 5  
b. 0.00592

E95. 0.90; See the explanation in the dialogue on pages 478–479 and in E23.

E97. This sentence is not correct: "To be more specific, the laws of probability say that if we were to conduct the same survey 100 times, asking people in each survey to rate the job Bill Clinton is doing as president, in 95 out of those 100 polls, we would find his rating to be between 47% and 53%.”
Chapter 9

Section 9.1

P1. a. (4.335, 5.704)
   b. The center is the same; too narrow; smaller.

P2. a. 2.262  
   b. 2.447  
   c. 3.106  
   d. 2.920  
   e. 2.704  
   f. 2.639

P3. a. (7.905, 46.095) (7.908 to 46.092 using Table B)
   b. (3.694 to 8.306)
   c. (−16.58, 34.577) (−16.572 to 34.572 using Table B)

P4. (4.23, 5.81); wider

P5. a. (97.969, 98.231) 
   b. You are 95% confident that the mean body temperature of all men is between 97.969°F and 98.231°F.
   c. No; the mean body temperature of all men is less than 98.6°F.

P6. a. No, but a sample of size 61 is large enough that this should not matter.
   b. (8.93, 11.59)
   c. You are 90% confident that the mean number of hours of study per week for all students taking this course is between 8.93 and 11.59. 90% confidence means that if you were to take 100 random samples of size 61 from the population of all students who take this course and calculate a confidence interval for each sample, you would expect 90 to contain the mean number of hours of study for all students.

P7. D

P8. a. false  
   b. false  
   c. false  
   d. false  
   e. true

P9. a. men: 0.397; women: 0.377 
   b. for the men, because the SD is larger than that for the women

E1. a. No; this is not a random sample but a group of volunteers.
   b. (50.69, 66.05)
   c. If this were a random sample of people, you would be 99% confident that if you blindfolded all people and had them attempt to walk the length of a football field, the mean distance walked before crossing a sideline would be between 50.69 yd and 66.05 yd.

E3. Conditions are met. You are 95% confident that the mean weight of the bottles produced that day is between 15.92 oz and 16.03 oz. Because 16 oz is one of the plausible values for the population mean, there is no need to adjust the machine.

E5. a. No. This is not a random sample of bags of fries.
   b. If this were a random sample, then you would be 95% confident that the mean mass of a small bag of fries at this McDonald's is between 69.926 g and 77.511 g (69.92 g and 77.52 g using Table B).
   c. No. The confidence interval contains 74 g, so 74 g is a plausible value for the mean mass of a small bag of fries.

E7. a. no, because there is no indication that this is a random sample of brochures 
   b. (8.895, 10.705)
   c. If these 30 could be considered a random sample of all brochures and if the readability found on the sampled page for each is typical of that brochure, you are 95% confident that if you tested the readability of all cancer brochures the average reading level would be between grades 8.895 and 10.705.

E9. a. third interval  
   b. first interval  
   c. third interval

E11. B

E13. a. increases  
   b. decreases  
   c. increases

E15. a. Skewed toward larger values; s is too small on average.
   b. s is smaller than σ = 112 in at least 61 of the 100 samples.
   c. The advertised 95% confidence intervals will be too narrow more often than they will be too wide when s is used as an estimate of σ. This means the actual capture rate will be less than 95%.

Section 9.2

P10. t = −2.9314

P11. t = 1.0853

P12. The lighter graph is the standard normal distribution because it has less area in the tails.
P13. a. One way is to use randNorm(100, 15, 7) and then calculate $t$.

P14. If the mean temperature at your desk is actually 72° F, the probability that temperatures taken on seven randomly selected days would give a value of $t$ greater than 2.9314 or less than $-2.9314$ is about 0.0262.

P15. If the mean Pell Grant for Minnesota college students is still $2178.82, the probability that a random sample of 35 students would have a $t$-statistic greater in absolute value than 1.0853 is about 0.285.

P16. a. $H_0$: The mean temperature at your desk, $\mu$, is 72°F. $H_a$: $\mu \neq 72°F$.
   b. yes; yes; no

P17. a. $H_0$: The mean Pell Grant in Minnesota, $\mu$, is $2178.82. H_a$: $\mu \neq 2178.82$.
   b. no; no; no

P18. a. $H_0$: The mean SAT score of State University students, $\mu$, is 1700. $H_a$: $\mu \neq 1700$.
   b. yes; no; no

P19. This might not be a random sample, but other conditions are met. Because the $t$-value, 2.918, exceeds the critical value, $t^* = 2.262$, reject $H_0$. The data suggest that the mean aldrin level differs from 4 nanograms.

P20. For males, with $t = -4.10$ and $P$-value 0.003, reject $H_0$: $\mu = 98.6$, where $\mu$ is the mean body temperature of all men in the population under study. For females, with $t = -0.48$ and $P$-value 0.64, do not reject $H_0$.

P21. a. Type I error b. Type I error c. Type I error

P22. a. $n = 16$ b. $\alpha = 0.05$

P23. a. one-sided (right-tailed) b. $x$ is the mean selling price for a sample of houses sold this month, and $\mu$ is the mean selling price of all houses for this month. c. $H_0$: $\mu$ is the same as the mean for last month. $H_a$: $\mu$ is greater.

P24. Conditions are met. Because the observed $t$-statistic, $-2.931$, is farther out in the tail than the critical value, $t^* = -1.943$, reject $H_0$. The data suggest that the mean temperature at your desk is less than 72°F. Alternatively, if the mean temperature at your desk is indeed 72°F, there is only a 1.3% chance of getting a result from a sample of size 7 as extreme as or more extreme than the result from your sample. So there is sufficient evidence to support the claim that the mean temperature at your desk is less than 72°F.

E17. Conditions are met. If the mean bottle weight were 16 oz, a $t$-statistic at least as large in absolute value as 1.022 would occur about one-third of the time. This means you do not have sufficient evidence to reject $H_0$. The data suggest that it is plausible that the mean amount of water in the bottles is 16 oz.

E19. This probably is not a random sample of students, but other conditions are met. Reject $H_0$. If it were true that the mean average error is 0, a $t$-statistic as extreme as or more extreme than 3.943 in absolute value would occur only 0.15% of the time.

E21. a. one-sided (left-tailed) b. $\bar{x}$ is the mean freezing point for a sample of ten bowls of salt water; and $\mu$ is the freezing point for salt water with this degree of salinity. c. $H_0$: $\mu = 32°F; H_a$: $\mu < 32°F$

E23. Conditions are met. If it is true that the mean weight of Munchie's Potato Chips is 10 oz, the probability of getting a $t$-statistic less than 2.55 is only 0.012. Reject $H_0$, because there is sufficient evidence that the mean weight of a bag of Munchie's Potato Chips is less than 10 oz.

E25. The chance of getting a random sample of young women with a $t$-statistic of 8.73 or larger is $2.7 \times 10^{-6}$. The difference in the mean height of the players and the mean height of all young women cannot reasonably be attributed to chance variation.

E27. Because the $P$-value, 0.333, is greater than $\alpha = 0.10$, do not reject $H_0$. The data suggest that it is plausible that the mean amount of water in the bottles is 16 oz.

E29. a. Because the $P$-value, 0.012, is less than $\alpha = 0.05$, reject $H_0$. There is sufficient evidence that the mean weight of a bag of Munchie's Potato Chips is less than 10 oz. b. No error was made. c. The 95% confidence interval for the mean weight is (9.73, 9.98), which suggests that the mean weight is less than 10 oz.

E31. D

E33. a. The $P$-value is 2(0.057).
b. The one-sided \( P \)-value is half as large as the two-sided \( P \)-value.

c. two-sided test

E35. Both have probability 0.05.

E37. a. \( \alpha = 0.10 \)  
b. \( n = 45 \)

c. one-sided test (assuming the alternative hypothesis is in the correct direction)

E39. a. They are both mound-shaped and centered at about 0. A has a smaller spread.  
b. B, because it has the larger spread

Section 9.3

P25. A. II  B. III  C. IV  D. I

P26. a. Do the analysis with and without the outlier.  
b. This isn't a random sample of bags of fries, but other conditions are met. In both cases, the \( P \)-value is very high (0.71 with the outlier and 0.87 without). Do not reject \( H_0 \). Neither analysis provides statistically significant evidence that the mean number of fries is different from 90.

E41. a. IV  b. III  c. I  d. II

E43. a. yes  
b. You are 90\% confident that the mean LOS for Insurer A is in the interval (2.22, 2.42).

c. no for \( n = 40 \), but yes for \( n = 4 \)

E45. a. No; do a log transformation.

c. No. Still more outliers are created, which is typical of strongly right-skewed data.

E47. a. No. Still more outliers are created, which is typical of strongly right-skewed data.

b. (102, 687); (69, 229)

c. Without outliers, the center is much lower and the width is much smaller. Confidence intervals apparently are quite variable when the distribution is highly skewed.

d. You are 95\% confident that the mean of the natural log of the weights of mammals' brains is between 2.33 and 3.63.

E49. a. transform  
b. On average, it takes 0.081 year to get an inch of rainfall; not quite, because there are still two outliers and a slight skew to the right.

c. much better

d. no, because a sample of size 128 is large enough that the sampling distribution of \( \bar{x} \) would be approximately normal

E51. a. The stemplots show marked skewness for the \( P/H \) data but near symmetry for the \( H/P \) data. There are outliers in both distributions.  
b. The interval (1.465, 3.674) contains the set of plausible values for the mean number of people per housing unit in the population of Florida counties.

c. You are 95\% confident that the true mean number of housing units per person by county is between 0.355 and 0.559.

d. (1.789, 2.817), which is quite different from the confidence interval in part b

e. housing units per person

Section 9.4

P27. a. yes  
b. (−12.760, 15.899)

c. the difference between the mean distance all left-handed volunteers will walk before crossing a sideline and the mean distance all right-handed volunteers will walk before crossing a sideline

d. no, because 0 is in the confidence interval

P28. Conditions are met except for the outliers on the high end. Including them, you are 90\% confident that \( -0.32 < \mu_{\text{female}} - \mu_{\text{male}} < 5.79 \). Excluding them, the interval is \( 1.27 < \mu_{\text{female}} - \mu_{\text{male}} < 5.75 \) and no longer contains 0. Because the results are so different, it would be good to gather more data to resolve the issue.

P29. Conditions are met except for the outliers. With the outliers, the \( P \)-value is 0.1392; without the outliers, the \( P \)-value is 0.012. Because the results are so different, it would be good to get more data.

P30. a. Results should be viewed with some caution because of the outlier. The \( P \)-value, 0.0568, is a bit larger than \( \alpha = 0.05 \), so this is not quite sufficient evidence to reject the hypothesis that any observed difference in means is the result of the randomization of the assignment of treatments to the babies.

b. The \( P \)-value is 0.4038, so there isn't sufficient evidence to reject the hypothesis of equal means for the exercise control and weekly report groups. An observed difference in
means of this size could quite reasonably be due to the random assignment of treatments to the babies.

c. The special exercises group is very close to having a statistically significantly lower mean than the weekly report group. The next strongest comparison is between the special exercises and exercise control groups, and the weakest is between the exercise control and weekly report groups.

P31. The measurements from the bottom have greater variability, so get most of the new measurements at the bottom to give your test greater power.

P32. a. The plausible values for the true difference in mean body temperatures, $\mu_{\text{female}} - \mu_{\text{male}}$, lie in the interval $(-1.149, -0.131)$. Or $(-1.038, -0.242)$ if you use the summary statistics.

b. Because this interval is entirely below 0, conclude that there is a statistically significant difference between the mean body temperature of men and of women.

c. one interval for the difference

E53. a. yes, but with some caution because of the high value for short days

b. (0.50, 9.46)

c. the difference between the mean enzyme concentration of all eight hamsters had they all been raised in short days and the mean concentration had they all been raised in long days

d. yes, because 0 is not in the confidence interval

E55. a. Conditions are met. A 95% confidence interval for the difference in initial mean pain levels is $(-7.092, 10.092)$, which includes 0. There isn't statistically significant evidence that the randomization failed to yield groups with comparable means.

b. You are 95% confident that the difference between the mean body mass indices of the two groups is in the interval $(-1.587, 2.5867)$, which includes 0. There isn't statistically significant evidence that the randomization failed to yield groups with comparable means.

E57. a. Use absolute values.

b. This amount of skewness should be okay for inference. If there were no difference in the means of the absolute values of the errors of the vertical line group and the horizontal line group, a $t$-statistic at least as large as 2.09 would occur only about 2.4% of the time. With a $P$-value this low, reject $H_0$. There is sufficient evidence to support the claim that students marking the midpoint on vertical line segments have a larger mean absolute error than the students marking the midpoint on horizontal line segments.

E59. a. No. No randomization is mentioned. Groups of size 8 are not large enough to compensate for this degree of skewness.

b. If the mean number of termites remaining had all dishes received a dose of 5 mg of resin is equal to the mean number remaining had all dishes received a dose of 10 mg of resin, a test statistic at least as extreme as $\pm 2.153$ would occur in 5.89% of the possible randomizations. You don't have statistically significant evidence that the size of the dose makes a difference in the mean number of termites remaining.

c. The main concerns are the lack of randomization and the skewness of the distribution of termites remaining in the 10 mg group. Repeat the experiment, randomly assigning treatments to the dishes and using a larger number of experimental units.

E61. With 95% confidence, $-4.71 < \mu_{\text{men}} - \mu_{\text{women}} < 9.94$, where $\mu_{\text{men}}$ is the mean heart rate for all men in this population and $\mu_{\text{women}}$ is the mean heart rate for all women in this population. Because 0 is in the confidence interval, there is insufficient evidence to say that the mean heart rate for men differs from the mean heart rate for women. Alternatively, the $P$-value is 0.4663, so there isn't statistically significant evidence that the mean heart rates differ. If the outlier among the men's heart rates is dropped, the 95% confidence interval for the difference in means becomes $-2.76 < \mu_{\text{men}} - \mu_{\text{women}} < 10.94$, and the conclusion remains the same.

E63. a. Type I

b. Decrease the level of significance.

c. There was no overlap in the concentrations for the two groups.
E65.  a. more like a survey of people’s ability to walk straight while blindfolded

b. If these were independent random samples and there was no difference between the mean distance men and women could walk, a t-statistic greater than 2.774 or less than −2.774 would occur with probability 0.0098. Reject $H_0$. There is strong evidence that there is a difference between these mean distances.

c. The lurking variable is the person’s height. Taller people tend to walk farther and men tend to be taller than women, so gender or height could explain the difference.

E67.  a. The problem does not say whether these are independent random samples. A test statistic of only $t = 1.50$ and a $P$-value of 0.0825 do not allow rejection of $H_0$, that coaching improves SAT math scores by 25 points.

b. Using a one-sample $t$-test to compare the average gain, 73, with the expected mean gain, 25 + 13 or 38, yields a $P$-value of 0.0074. Reject $H_0$ and conclude that coaching improves the average gain by more than the expected 38 points.

Section 9.5

P33.  a. repeated measures; completely randomized

b. Conditions are met. With a $P$-value of 0.005, reject $H_0$. There is sufficient evidence to conclude that standing increases the mean pulse rate by more than would be expected by chance for this group of subjects.

c. Conditions are met. With a $P$-value of 0.0023, reject $H_0$. There is sufficient evidence to conclude that standing does increase the mean pulse rate by more than would be expected by chance.

d. Conditions are met. With a $P$-value less than 0.0001, reject $H_0$. Once again, there is sufficient evidence to conclude that standing does increase the mean pulse rate by more than would be expected by chance. The evidence is even stronger than in the matched pairs design.

e. yes

P34.  a. independent samples

b. log transformation

c. Conditions are met. With a $P$-value of 0.0248, reject $H_0$. There is sufficient evidence to conclude that the mean brain weights of species of birds and species of fish differ.

P35.  a. Conditions could not be randomly assigned.

b. With a $P$-value of 0.33, the difference in lung clearance between the twins can reasonably be attributed to chance alone.

c. Independently, take random samples from rural and urban populations.

d. The analysis of differences reduces the person-to-person variation.

e. You would not have to find a criterion on which to pair the subjects.

P36. No; statistics isn’t needed to know that hens’ eggs are longer than they are wide.

P37.  a. No; there is no random assignment into treatment groups.

b. It is better, but there still is no random assignment of treatments.

c. Randomly divide a group of students into the two treatment groups.

e. You are 95% confident that the mean difference between the distances of bears launched using four books and the distances of bears launched using one book is between −23.07 in. and 146.74 in. Because 0 is in the interval, there isn’t sufficient evidence to support the claim that launch angle makes a difference in how far gummy bears soar.

d. no, because the sets of distances are not independent samples

e. yes

P38.  a. one-sample test of the mean of the differences; one-sided test

b. Conditions are met. Because the $P$-value, 0.0016, is less than $\alpha = 0.01$, reject $H_0$. If these groups can be considered a random sample of all possible teams of students, you have statistically significant evidence that the mean distance on the tenth launch is greater than the mean distance on the first launch.

c. yes, because teams should improve with practice; Type I error

d. no, because the sets of distances are not independent samples

e. The 95% confidence interval for the difference $\mu_{\text{bottom-mid-depth}}$ is (0.5655, 1.4145). There is statistically significant evidence that
the mean difference is not 0, if the conditions for inference are met.

b. no
c. one-sample test of the mean difference

E75. a. two-sample test
b. Conditions are met. If the higher step did not increase the mean gain in heart rate, a t-statistic larger than 2.778 would occur with probability 0.0051. This is strong evidence against H₀. There is statistically significant evidence that using the higher step increases heart rate more than does using the lower step.

e77. a. $H_0: \mu_d = 0$; $H_1: \mu_d > 0$, where $\mu_d$ is the mean difference in the number of stings
b. It is unclear whether any randomization took place. The distribution of the differences is fairly symmetric with the exception of one outlier. Do the analysis both with and without the outlier.
c. With a P-value of 0.0542 or 0.0528 (without the outlier), don’t reject $H_0$. There isn’t statistically significant evidence that there would have been more new stingers in the previously stung cotton balls than in the fresh cotton balls had Free been able to give all cotton balls both treatments.

E79. a. two independent samples
b. Pair rooms, one with carpet and one without, that are near each other and have similar uses. Alternatively, use a repeated measures design, giving each room both treatments and randomizing which eight rooms get the carpet first.

E81. It is not the center but the variability that is of interest.

E83. a. two independent samples
b. Pair students on their ability to estimate, and then randomly assign one of each pair to feet and one to meters. Or have each student estimate in both feet and meters.
c. Matching students on their ability to estimate distances would be difficult. Repeated measures might not work, because the student might change one estimate to the other by multiplying by a constant.
d. Because one mean is in feet and the other is in meters, of course the means aren’t going to be equal. Further, the question asked was whether the students were as accurate in meters as they were in feet, and these data can’t answer that question because you don’t know the true length.

E85. a. yes, if eruptions can be selected at random
b. unpaired

E87. a. $\bar{x} - \mu_{A} < 12.82$; $0.06 < \mu_{B} - \mu_{A} < 33.94$; the first estimates the mean difference for all accounts, and the second estimates the mean difference for all accounts with nonzero differences. The latter is the more reliable estimate because the conditions are better met.
b. You would expect about $\frac{15}{46}(1000)$, or 375, of the 1000 accounts to show a difference between the book and audit values. Multiplying both ends of the interval by 375 gives a confidence interval for the total, (22.5, 12,727.5).

Chapter Summary

E89. a. yes
b. You are 95% confident that if you were to ask all highly trained athletes how long they sleep, the mean number would be between 5.63 h and 7.71 h.

E91. Conditions are met. If it is true that the mean distance had all students used four books is the same as the mean distance had all students used one book, a t-statistic greater than 3.65 or less than $-3.65$ would occur with probability 0.0005. Reject $H_0$. There is statistically significant evidence that the number of books changes the mean distance of gummy bear flight.

E93. Conditions are met. You are 95% confident that the mean difference between the life expectancies of females and males is in the interval (1.493, 5.147). There is sufficient evidence to say that the difference between the means is positive, which implies that the population mean life expectancy for females is larger than that for males. (Using a test gives a P-value of 0.0026.)

E95. D

E97. a. Transform the salaries.
b. Proceed if the firm has more than 500 people.

E99. a. yes, because the chance of getting a randomly selected team of 16 players with a batting average of .276 or higher is only 0.0188 (Use $z = \frac{x - \mu}{\sigma/\sqrt{n}}$ because \sigma is known to be .050.)
b. no, because the chance of getting a randomly selected team of 16 players with a batting average of .256 or higher is 0.3156

Chapter 10

Section 10.1

P1. a. 12.5  b. $\chi^2 = 3.28$

P2. 0

P3. P-value is small (about 146/5000 or 0.0292). Reject the $H_0$ that the die is fair.

P4. P-value is near 0. Reject the $H_0$ that the die is fair.

P5. a. $\chi^2$ with $df = 3$ is 7.81. Reject the $H_0$ that the die is fair.
   b. $\chi^2$ with $df = 19$ is 30.14. Do not reject the $H_0$ that the die is fair.

P6. a. 0.9194  b. 0.0304

P7. Conditions are not met, as this is a convenience sample. With $\chi^2 = 1.25$ and $df = 3$, the P-value is 0.7410. Do not reject the $H_0$ that the blood types O:A:B:AB occur in the ratio 9:8:2:1.

P8. a. Conditions are not met, as this is an observational study. The expected count in each category is 117.5. With $\chi^2 = 16.723$ and $df = 3$, the P-value is 0.0008. Reject the $H_0$ that seizures are equally likely during these four phases of the moon. The largest deviation occurs during the last quarter. There were fewer seizures during the full moon than the model predicts.
   b. Were the quarters of equal length? Over how many months were these data collected?

P9. a. Conditions may not be met because your class probably will not be large enough to have expected frequencies of 5 in each category.
   b. Combining the last three categories into 2 or more, with probability 0.275, $\chi^2 = 44.20$ with $df = 2$ gives a P-value of 2.5 · $10^{-10}$. Reject the $H_0$ that the distribution of the number of children in the families of statistics students is consistent with the distribution for families nationwide.

P10. a. Conditions are met. With $z = 2.80$, the P-value for a two-tailed test is about 0.005. Reject the $H_0$ that spinning a coin is fair.
   b. Conditions are met. With $\chi^2 = 7.84$ and $df = 1$, the P-value is about 0.005. Reject the $H_0$ that spinning a coin is fair.
   c. The P-values are equal and $z^2 = \chi^2$.

P11. a. Conditions are met. With $z = -5.4$, the P-value for a two-tailed test is 6.7 · $10^{-8}$. Reject the $H_0$ that spinning a checker with putty on the crown side is fair.
   b. Conditions are met. With $\chi^2 = 29.16$ and $df = 1$, the P-value is 6.7 · $10^{-8}$. Reject the $H_0$ that spinning a checker with putty on the crown side is fair.
   c. The P-values are equal and $z^2 = \chi^2$.

E1. Conditions are met. With $\chi^2 = 94.189$ and $df = 5$, the P-value is 8.84 · $10^{-15}$. Reject the $H_0$ that the die is fair.

E3. Conditions are not met, as this is probably an observational study. With time periods of 15, 46, 120, and 184 days, $\chi^2 = 1.39$, $df = 3$, P-value = 0.7080. Do not reject the $H_0$ that the number of people admitted in each time period is proportional to the length of that time period.

E5. With $\chi^2 = 0.47$ and $df = 3$, the probability of a chi-square statistic less than or equal to 0.47 is 0.0746. The data fit the model so well that it casts some doubt on the conduct of the experiment.

E7. Crossley: With $\chi^2 = 56.67$ and $df = 3$, the P-value is about 3 · $10^{-12}$.
   Gallup: With $\chi^2 = 108.27$ and $df = 3$, the P-value is about 3 · $10^{-21}$.
   Roper: With $\chi^2 = 229.07$ and $df = 3$, the P-value is about 2 · $10^{-49}$.
   None of the three polls are close to fitting the actual results, but Crossley fits best.

E9. Conditions are not met, as these are the populations of fatalities for the selected days and hours.
   a. With $\chi^2 = 47.10$ and $df = 1$, the P-value is 6.7 · $10^{-12}$. Reject the $H_0$ that the difference between the number of traffic fatalities during the first four hours after the Superbowl and (half) the number during those hours on the previous and the following Sundays can be attributed to chance variation.
   b. With $\chi^2 = 0.36$ and $df = 1$, the P-value is 0.5485. Do not reject the $H_0$ that the number involved in traffic accidents fits the 1:2 model.

Brief Answers to Selected Problems  871
The observed difference in fatalities could reasonably be attributed to chance variation alone.

E11. Answers will vary.

E13. a. As the sample size increases, the values of sample statistics such as proportions tend to get closer to the true population values, thus improving the power to reject a false $H_0$.

b. Yes, if $H_0$ is false, all else being equal, the $\chi^2$ statistic tends to be larger with larger $n$ than with smaller $n$, giving you more power to detect the false $H_0$. Also, the sampling distribution of $\chi^2$ becomes closer to the continuous chi-square distribution.

E15. a. Each digit should occur $\frac{1}{10}$ of the time.

b. The rivers are not randomly selected, but the final digit of the lengths can be regarded as somewhat random from the population of final digits of measured rivers. With $\chi^2 = 131.77$ and $df = 9$, the $P$-value is $5.13 \times 10^{-24}$. Reject the $H_0$ that the final digits are equally likely.

Section 10.2

P12. b. It appears that females are more likely than males to prefer a bath.

P13. a. marginal  b. conditional  
c. marginal  d. neither

P14. For both males and females, expected frequencies are 13.5 and 13.5 for bath and 36.5 and 36.5 for shower.

P15. b. $\frac{40}{360} = 0.111$; $\frac{100}{360} = 0.278$; $\frac{60}{360} = 0.167$; $\frac{50}{360} = 0.139$; $\frac{110}{360} = 0.306$

c. first row: 11.33, 22.00, and 6.67

d. No. Population III appears to have too many members that fall into Categories A and B and too few in the others.

P16. $\chi^2 = 11.42$; $df = 1$; If male and female students have the same preferences, the probability of getting a test statistic of 11.42 or larger is less than 0.001 (or 0.0007).

P17. $\chi^2 = 64.84$; $df = 8$; yes

P18. $\frac{3080}{5000} \cdot 1000 = 616$; $\frac{1770}{5000} \cdot 1000 = 354$; $\frac{150}{5000} \cdot 1000 = 30$

P19. The $P$-value is between 0.15 and 0.20 (or 0.1826). Don’t reject the $H_0$ that the proportion of people in each category is the same for each country.

P20. Yes. With $\chi^2 = 4.82$ and $df = 1$, the $P$-value is between 0.025 and 0.05 (or 0.0281).

P21. With $\chi^2 = 11.11$ and $df = 2$, the $P$-value is between 0.0025 and 0.005 (or 0.0039). Reject the $H_0$ that if both treatments could be assigned to all subjects, the distributions of outcomes would be the same. The nitric oxide treatment resulted in more Not severe hemorrhages and fewer Severe ones.

E17. a. The observed frequencies for 2003 are 190, 200, 90, and 520.

b. the populations of adult residents of the United States in the respective years

c. Conditions are met. With $\chi^2 = 113.4$ and $df = 9$, the $P$-value is almost 0. Reject the $H_0$ that the distributions of responses in the populations are the same for each year.

E19. b. $z = 2.915$; If all subjects could have been assigned each treatment and which treatment the subject received made no difference in survival, the probability of getting a difference in sample proportions as large in absolute value as from the actual experiment is only 0.0036.

c. $\chi^2 = 8.495$; If all subjects could have been assigned each treatment and which treatment the subject received made no difference in survival, the probability of getting a $\chi^2$ statistic as large as or larger than from the actual experiment is only 0.0036.

d. The $P$-values are equal and $z^2 = \chi^2$.

E21. Conditions are met. With $\chi^2 = 34.575$ and $df = 4$, the $P$-value is close to 0. Reject the $H_0$ that the three treatments would give the same distributions if all children could have received each treatment.

E23. Conditions are met. With $\chi^2 = 17.57$ and $df = 6$, the $P$-value is between 0.005 and 0.01 (or 0.0074). Reject the $H_0$ that the distributions of responses in the population are the same for these years.

E25. Conditions are met. Reject the $H_0$. A value of $\chi^2$ as large as or larger than 18.19 (or a $P$-value as small as 0.0011) is extremely unlikely to occur if the proportion of the population in each category is the same for each country. The United States appears to have less violent crime and a larger proportion of non-victimized households.
E27. Conditions are met. Reject the $H_0$. A value of $\chi^2$ as large as or larger than 271.56 (or a $P$-value close to zero) is extremely unlikely to occur if the distribution of responses throughout the populations is the same for each country. The United States appears to have a smaller proportion of individuals who report drinking every day, a few times a week, or about once a week than either Canada or Great Britain, and a larger proportion who report never drinking.

E29. a. No. Some Internet users will fall into more than one category and some into neither.
b. Yes. If the proportion of the population that uses the Internet for sending and reading e-mail hasn’t changed, the probability of getting a test statistic as large as or even larger than that from this test is 0.0914.

E31. One meaning of “homogeneous” is “essentially alike.” Homogeneous populations would be populations that are alike in the characteristics you are observing.

Section 10.3

P22. a. not independent  b. not independent  
c. depends on recent fashion trends  
d. not independent  e. independent  
f. not independent

P23. A and C

P24. Expected frequencies for right-handed are 955.86 and 1048.14.

P25. Conditions are met. Reject $H_0$. If gender and handedness are independent, the probability of getting a value of $\chi^2$ as large as or larger than 11.81 from a random sample of size 2237 is only 0.0027. It appears that women are more likely to be right-handed and men are more likely to be left-handed or ambidextrous.

P26. a. test of independence  
b. Yes, with $\chi^2$ of 21.09, the $P$-value is about 0.0001 so there is evidence that gender is associated with effort at eating a healthy diet.  
c. Yes, but the evidence is much weaker. (Depending on rounding, $\chi^2$ is cut in half to about 10 or 11 and the $P$-value is 0.012 to 0.019.)

P27. If the sample is random, the conditions are met. Reject the $H_0$. If, among all students at this middle school, having breakfast three or more times a week and thinking they have a healthy diet are independent, the probability of getting a chi-square of 4.475 or higher from a random sample is 0.0344.

P28. a. With $\chi^2 = 18.37$, the $P$-value is 0.000018. There is a statistically significant association between gender and attainment of a high school education. You have strong evidence of a weak association.
b. With $\chi^2 = 475.13$, the $P$-value is effectively 0. There is a highly statistically significant association between gender and attainment of a high school education. The association is stronger than in part a.
c. In part a, the difference is too small to be of much practical significance. In part b, the difference is larger and more meaningful.

E33. a. In some regions (such as Alaska and parts of the Southwest), adults are younger than in other regions and so would tend to have children in lower grade levels.
b. for K–8: NE 0.1159; MW 0.1525; South 0.2519; West 0.1697
c. for K–8: NE 5,630,698; MW 7,407,049; South 12,233,362; West 8,244,951

E35. a. Conditions are met. Don’t reject the $H_0$. If a random sample this size is taken from a student population where class year and favorite team sport are independent, there is a 30.4% chance of getting a $\chi^2$ value of 10.596 or larger.
b. no, as you haven’t rejected the $H_0$
c. Take a random sample of, say, 80 freshmen and ask them their favorite team sport. Repeat with sophomores, juniors, and seniors.

E37. b. With $\chi^2 = 456.9$ and a $P$-value close to 0, reject the $H_0$ that the difference in the proportions of males and females who survived can reasonably be attributed to chance. These are not the results you would expect if people were placed on lifeboats without regard to gender.

E39. a. observational study  
b. homogeneity  
c. You do not have a random sample of Sundays, or even of Super Bowl Sundays. If the probability that a person in an alcohol-related
accident is killed is the same the first four hours after a Super Bowl as it is during the same four hours on a Sunday before or after a Super Bowl, a $\chi^2$ as high as or higher than 138.69 is extremely unlikely. Reject the $H_0$. Something other than random variation must account for the high incidence of fatalities following a Super Bowl.

Chapter Summary

E41. Conditions are met for a test of homogeneity. With $\chi^2 \approx 66.03$, the $P$-value is about 0.000015. Reject the $H_0$ that if in every year from 2001 through 2005 and in 1997, you had asked all adults in the United States this question, the distribution of responses would be the same for each year. The most significant trends are that fewer people have no opinion and more people think global warming has already begun.

E43. b. The infirmary was used most heavily on Monday and Thursday. There were relatively few admissions on Friday. Perhaps students saved their problems until Thursday after classes were over and perhaps the infirmary wasn't open weekends, so students who got sick on the weekend had to wait until Monday.

c. Yes for test of independence, as there was one large population that was classified according to day and severity of problem; no, as this isn't a random sample

d. Because the table includes all students who visited the infirmary during a school year, these data cannot reasonably be considered a random sample taken from one large population. With $\chi^2 \approx 11.023$ and $P$-value 0.527, don't reject the $H_0$. These are typical of the results you would see if there were no association between severity of problem and day of the week.

e. You might group Monday, Tuesday, and Wednesday together and Thursday and Friday together.

f. With $\chi^2 \approx 3.463$, and a $P$-value of 0.326, the association is not statistically significant here either. The differences in the proportions of the various problems reported on weekdays when students attend class and on weekdays when students don't attend class can reasonably be attributed to chance.

E45. With $\chi^2 \approx 0.88$ and a $P$-value of 0.644, there isn't sufficient evidence to reject the percentages claimed.

E47. a. Homogeneity

b. Conditions are met. Don't reject the $H_0$. You can attribute the differences to the fact that you have only a sample of adults from each educational level and not the entire adult population. A value of $\chi^2$ of 3.202 or even larger is quite likely to occur ($P$-value close to 0.20) for three samples of this size if the same proportion of adults in each population watched the Super Bowl.

E49. a. first row: 20, 16  c. no

E51. a. Yes, conditions are met for a test of goodness-of-fit. With $\chi^2 = 5.456$ and a $P$-value of 0.2436, don't reject the $H_0$. You don't have statistically significant evidence that the distribution of grades differs from that of 1996.

b. Because you know the population standard deviation for 1996, $\sigma = 1.30$, you can use that value in the $z$-statistic. With $z = 1.088$ and a $P$-value for a right-tailed test of 0.1383, don't reject the $H_0$. You don't have statistically significant evidence that the mean grade in 2001 increased over that of 1996.

E53. This sample cannot reasonably be considered a simple random sample taken from one large population. With $\chi^2 \approx 133.052$ and a $P$-value close to 0, reject the $H_0$. These are not results like those you would expect if people were placed on lifeboats without regard to class of travel. The explanation apparently is that first-class passengers were the first to be allowed into lifeboats and third-class passengers were last.

E55. a. B and I  b. A and III  c. C and II

Chapter 11

Section 11.1

P1. a. 9 calories per gram of fat  
b. the number of calories in a serving containing no grams of fat  
c. Calories in pizza come from carbohydrates and protein as well as from fat; there
may be other sources of variation in the measurement process.

P2. a. \( \mu_y = 0 + x \), where \( x \) is the arm span and \( \mu_y \) is the mean height at that arm span.
b. \( \hat{y} = 7.915 + 0.952x \)
The height increases, on average, by about 0.952 inch for each 1-inch increase in arm span.

P3. a. \( \hat{y} = b_0 + b_1x \)

b. \( \text{opening day} = 97.37 + 1.75 \cdot \text{swe} \)
c. If \( \text{swe} = 31.0, \text{opening day} = 151.62 \) or 152 (June 1).
d. The random variation, \( \epsilon \), should be relatively large, given all the conditions that can affect the opening date.

P4. 0.341366

P5. a. 5; greater variability in the conditional distributions of \( y \) results in a larger value in the numerator of \( s_{b_1} \).
b. 3; a smaller spread in the values of \( x \) results in a smaller value in the denominator of \( s_{b_1} \).
c. 10; all else being equal, a larger sample size tends to result in less variability in the estimates of parameters.
d. The theoretical slope does not matter, all else being equal.
e. The theoretical intercept does not matter, all else being equal.

P6. a. 0.1472

b. \( \hat{y} = 1.71525 + 0.52490x \)
c. The first value is the slope of the regression line, \( b_1 \). The second value is the estimated standard error of that estimated slope, as calculated in part a.

P7. a. \( y = \beta_0 + \beta_1x = 10 + 2x; \)
\( \hat{y} = b_0 + b_1x = 9.91 + 2.03x \)
b. \( s = 3.127, \) close to the theoretical value of 3
c. \( s_{b_1} = 0.3496 \) is close to 0.335.
d. yes

P8. V, II, I, IV, III

P9. I, III, II, IV.

E1. a. Yes for \( (hp, price) \), but note there is a tendency for the spread in \( y \) to increase as \( x \) increases. In the plot of \( (hp, mpg) \), there is a fairly strong negative trend with a hint of curvature. Yes for \( (l, mpg) \), except for two potentially influential points on the left.
No for \( (rpm, mpg) \), because the variation in responses is so great at the larger values of \( x \) that no linear trend is apparent.
b. \( b_1 = -0.069. \) For every increase of 1 unit in horsepower, the gas mileage tends to drop by 0.069 mile per gallon. \( s_{b_1} = 0.0233. \)
c. The estimated slope appears under “Parameter Estimates” in row “HP” and column “Estimate.” The standard error is found in column “Std Error” and row “HP.” The value of \( s \) is the Root Mean Square Error \( = 4.7785. \)
\( s \) is the estimate of the common variability in the miles per gallon for each fixed horsepower.

E3. a. \( \pi \), or about 3.1416. For every 1 cm increase in the distance across, the circumference will increase by 3.1416 cm.
b. The linear model that fits the situation perfectly is \( C = \pi d \). However, it is difficult to measure \( C \) and \( d \) with great accuracy.

E5. a. The soil samples should have the larger variability in the slope because the distance of \( y \) from the regression line tends to be larger compared to the spread in \( x \).
b. \( s_{b_1} = 0.3617 \). As predicted, the standard error for the soil samples is much larger.

E7. a. The plot shows a linear trend with large but homogeneous variation in heights across values of age.
b. The plot shows a linear trend with homogeneous variation that is smaller than the variation in part a.
c. The plot shows a linear trend with variability in height increasing with age.
E9. Choose a value for $x$.

Predict the mean value of $y$ by finding $\hat{y}$ through the regression line.

Generate a value at random from a normal distribution with mean 0 and standard deviation approximately equal to $s = 0.3414$.

Add that value to the $\hat{y}$ calculated above. This is your estimate of the redness for a rock with sulfate percentage $x$.

Complete the above four steps for each value of $x$ in the study.

Fit a least squares regression line through the resulting points.

Use the regression line to predict the redness when $x$ is 2.72, the value for Half Dome.

This model is reasonable given the limited information we have, but it may not be reasonable given that this is a sample of only a few rocks and we can’t be sure that (1) the relationship is linear, (2) the variation in $y$ is constant for all values of $x$, or (3) the distribution of the errors is normal.

Section 11.2

P10. $t = 0.3688$. With $df = 4$, the $P$-value is greater than $2(0.25) = 0.5$ from Table B and 0.731 from a calculator. Do not reject the $H_0$ that the true slope is 0.

P11. Small (in absolute value); there is no pronounced linear trend and the variation in the residuals is large.

P12. $t = 5.6982$. With $11 - 2 = 9$ degrees of freedom, the $P$-value from the calculator is 0.0003. Reject the hypothesis that there is no linear relationship between percentage of sulfur and redness.

P13. a. $\hat{y} = 25.232 + 3.291x$. The temperature rises, on average, about 3.291 degrees if the number of chirps per second increases by 1.

b. The residual plot shows no obvious pattern and little heterogeneity, so a linear model fits the data well. The dot plot of the residuals shows no outliers or any other indications of non-normality.

c. The $P$-value is close to 0. Reject the $H_0$ that there is no linear relationship between rate of chirping and temperature.

P14. Conditions are not met because this is not a random sample from any well-defined population. The plot of these data does not show a linear relationship. The residuals have a slightly skewed distribution. Proceeding anyway, the regression equation is $IQ = 0.9969 \cdot \text{head circumference} + 45.050$.

With a test statistic $t = 0.5904$ and 18 degrees of freedom, the $P$-value is 0.5622. Do not reject the $H_0$. There is no statistically significant evidence of a linear relationship between $\text{head circumference}$ and $IQ$.

P15. a. You are 90% confident that the slope of the true regression line for predicting titanium dioxide content from silicon dioxide content for Mars rocks is in the interval $-0.0429$ to $-0.0245$.

b. $s = 0.0257$ is an estimate of $\sigma$, the standard deviation of the errors from the true regression line. It can also be interpreted as an estimate of the variability in the titanium dioxide percentage at each fixed silicon dioxide percentage.

P16. You are 90% confident that the slope of the true regression line for predicting titanium dioxide content from silicon dioxide content for Mars soil is in the interval $-0.0714$ to $0.1251$.

Because the confidence interval for the soil samples entirely overlaps the one for the rocks, you can’t conclude that the slopes are different.

E11. a. $s_{b_1} = 0.0233$

b. Yes. With $t = -2.98$ and associated $P$-value of 0.0106, reject the $H_0$ at the 5% level of significance. Conclude that the slope is significantly different from zero. Note, however, the slight curvature may mean that a linear model is inappropriate.

c. The 90% confidence interval estimate based on 13 degrees of freedom is $s_{b_1} (-0.1107, -0.0281)$. This interval gives the plausible values of the expected decrease in miles per gallon per 1 unit increase in horsepower.

E13. Conditions are met, except that the variation in $y$ tends to increase with $x$. With $t = 1.40$ and 13 degrees of freedom, the $P$-value is 0.185. Do not reject $H_0$; there is no evidence of a linear association between expected gas mileage and the maximum speed of the engine in rpm's.

E15. a. i. You are 95% confident that the true slope of the regression line for predicting the
temperature from the chirp rate is in the interval 1.99 to 4.59.

ii. You are 95% confident that the true slope of the regression line for predicting chirp rate from temperature is in the interval 0.128 to 0.296.

b. When you reverse the roles of chirp rate and temperature, the entire regression line changes. The unit of the slope changes from degrees per chirp to chirps per degree.

E17. To test that $\beta_1 = 1$, $t = -1.454$. With 13 degrees of freedom, the $P$-value is $2(0.08485) \approx 0.1697$. Do not reject the hypothesis that the slope is 1. Alternatively, a 95% confidence interval for the slope is $0.9517 \pm 0.0718$, which includes 1 as a plausible value for $\beta_1$.

E19. $\text{size} = 31.83 - 0.712 \text{ acid}$. For every increase of 1 $\mu g/ml$ in acid concentration, the radius of the fungus colony tends to decrease by 0.712 mm, on average.

Assuming the treatments were randomly assigned, the conditions are met. With $t = -19.84$ and 10 degrees of freedom, the $P$-value is essentially 0. Reject $H_0$: $\beta_1 = 0$. There is a negative linear relationship between colony radius and acid concentration.

E21. $\text{price} = -32.3263 + 39.4553 \text{ bedrooms}$

The statistically significant slope indicates that the expected increase in price is about $39,455 per additional bedroom. But there is only one two-bedroom house and there are three four-bedroom houses. To get a fairer picture of the true relationship between average selling price and number of bedrooms, more two- and four-bedroom houses should be sampled.

Section 11.3

P17. Conditions are met. $t = 10.7793$. With 9 degrees of freedom, the $P$-value is $1.9 \cdot 10^{-6}$, which is very close to 0. Reject the $H_0$ that the slope is 3.

P18. a. The untransformed data clearly show curvature. Both the log transformation and the log-log transformation straighten the pattern considerably. It is difficult to choose between them on the basis of the residual plots or dot plots of the residuals. However, because there tends to be a close-to-cubic relationship between length and weight, you might choose the log-log transformation.

b. Conditions are met, assuming this is a random sample. $t = -4.3230$. With 7 degrees of freedom and a two-tailed test, the $P$-value is 0.0035. Reject the $H_0$ that the slope of the true regression line for ln(weight) versus ln(length) is equal to 3.

c. You are 95% confident that the slope of the true regression line for (ln(length), ln(weight)) would be in the interval (2.244, 2.779). The 95% confidence interval for widemouth bass was (3.301, 3.460). Because there is no overlap in these confidence intervals, you can conclude that the power function that relates length to weight of a black crappie has a smaller exponent than that for widemouth bass.

P19. a. The distribution of log(img) should be approximately normal with mean around 1.654 and standard deviation around $s = 0.2588$. About 45.082.

b. You are 95% confident that in the population of all countries, the slope of the regression line for (fr, log(img)) is (0.1423, 0.2477). This indicates that the exponential function of the form $y = ab^x$ that best fits the relationship between fr and img has plausible values of $b$ between $10^{(0.1423)}$ and $10^{(0.2477)}$, or 1.388 to 1.769. Thus, the img increases from about 39% to 77% for every one-unit increase in fr.

P20. a. A linear regression model will not work well here as the scatterplot and the residual plot both show curvature.

b. Squaring the percent successful straightens the pattern rather well (a power model does almost as well): $\text{percent successful}^2 = -51.40 + 426.632 \cdot \text{number of chimps}$.

c. You are 95% confident that the true slope of the linear relationship between the square of the percent of success and the number of chimps in the hunting party is between 366.34 and 486.92.

E23. a. Yes. The residuals are more evenly scattered. The slope of the regression line went from negative to positive, but is still not significantly different from 0.

A log-log transformation allows the residuals to look more randomly scattered and their...
distribution to be more symmetric, but the slope is not significantly different from 0.

b. The regression line allows you to explore the relationship between crime rate and population. A significance test allows conclusions such as the following: The slope of the relationship between the crime rate and total population of the city is no different from what you would expect if the crime rate had been assigned at random to the cities.

**E25.**

a. The plot of number of police versus violent crime rate shows a bit of upward curvature and shows that variation in y is increasing with increasing x. The scatterplot of ln(police) versus (violent crime rate) shows a much more linear trend. The regression equation is \( \ln(\text{police}) = 1.42 + 0.00334(\text{violent crime rate}) \).

b. Exponentiating both endpoints of the interval \((0.00126, 0.00542)\) gives an interval estimate of \((1.0013, 1.0054)\) for the base of the corresponding exponential function. Thus, if City A has a crime rate 1 unit greater than City B, the number of police officers in City A is expected to be between 0.13% and 0.54% greater than the number in City B.

**E27.**

Using a log-log transformation to reduce curvature, the regression line is \( \ln(\text{flow}) = -1.01 + 1.12 \ln(\text{static}) \). The slope is significantly different from 0 with a \( P \)-value below 0.0001.

**E29.**

a. No. A log-log transformation will reduce the curvature.

b. The fitted equation is \( \ln(\text{distance}) = 3.7514 + 0.5152 \ln(\text{height}) \). Conditions for inference are not met, as this is not a randomized experiment, but the confidence interval is still useful. You are 95% confident that, for every unit increase in the natural log of the height, the natural log of the distance will increase, on average, by an amount between 0.4681 and 0.5623.

c. These data show much more curvature, which is not eliminated by the log-log transformation. A more complex model is needed. The shelf, by eliminating the initial downward velocity, clearly made a difference in the relationship between distance and height.

**E31.**

a. This is a random sample of countries. The trend is fairly linear, and the variation in residuals is quite uniform. The distribution of the residuals looks quite symmetric with three outliers. The outliers are a reason for some caution.

b. Reject the \( H_0 \) that these data came from a population where the slope of the true regression line for predicting fertility rate from life expectancy for women is 0.

c. No. This was not an experiment where conditions or treatments were randomly assigned.

d. The large area of white space in the middle of the graph suggests two clusters. Each of these clusters taken separately may tell a very different story.

**Chapter Summary**

**E33.**

a. I, III, IV, and V, although IV shows no linear trend and V has an influential data point

b. IV

c. V, but III is close

d. II: try a log-log transformation. V: remove the influential data point.

**E35.**

a. If one pizza has 1 g of fat less than another, you expect to have 4.996 to 9.530 fewer calories.

b. \((24.98, 47.65)\)

**E37.**

a. Conditions for inference are met. The regression equation is \( \text{length} = -1037.1 + 0.5771 \cdot \text{year} \). You are 95% confident that the slope of the true linear relationship between length of a film and the year in which it was released is between 0.1048 and 1.0494. Because 0 is not in the interval, you have statistically significant evidence that there is a positive linear relationship.

b. Conditions for inference are met. The regression equation is \( \text{rating} = 0.97 + 0.01365 \cdot \text{length} \). You are 95% confident that the slope of the true linear relationship between a film’s
rating and its length is between 0.00121 and 0.02609. Because this interval lies entirely above 0, you have statistically significant evidence of a positive linear relationship.

c. You would expect a positive linear relationship between year and rating, but the trend appears negative.

Conditions for inference are met. The regression equation is \( rating = 21.1 - 0.00959 \text{ year} \).

For testing the \( H_0 \) of zero slope, \( t = -1.3381 \) with 18 degrees of freedom, giving a \( P \)-value of 0.1975. Do not reject the \( H_0 \) of zero slope. You do not have statistically significant evidence that the slope of the relationship between rating and year for all movies is different from 0.

d. For the longer movies, conditions for inference are met. The regression equation is \( rating = 31.3 - 0.01458 \text{ year} \).

For testing the \( H_0 \) of zero slope, \( t = -1.86 \) with 8 degrees of freedom, giving a \( P \)-value = 0.100. This is stronger evidence of a negative linear relationship between a movie’s rating and the year in which it was released than when short and long movies were considered together.

For the shorter movies, conditions for inference are met. The regression equation is \( rating = 59.7 - 0.02954 \text{ year} \).

For testing the \( H_0 \) of zero slope, \( t = -3.059 \) with 8 degrees of freedom, giving a \( P \)-value = 0.016. This is much stronger evidence of a negative linear relationship between a movie’s rating and the year in which it was released than when short and long movies were considered together.

Ratings tend to go down over the years, but this overall decreasing trend gets masked by the fact that more of the longer movies, with higher ratings, are from recent years.

E39. If “standard deviation of the responses” is interpreted as meaning “for each fixed \( x \)” then both are descriptions of \( \sigma \) and so are equal. If “standard deviation of the responses” is interpreted to mean the SD of all values of \( y \), then that value will be larger than the SD of the prediction errors.

---

**Chapter 12**

**Section 12.1**

P1. most effective: Treatments 1 and 5, because they are centered at the lowest values; least effective: Treatments 3 and 4

P2. The centers vary widely, as do the spreads. As the mean increases, the spreads tend to decrease.

P3. The treatment with the larger mean tends to have an SD that is 0.16 cm smaller.

P4. Conditions are met. With \( t = 0.702 \) and \( P \)-value = 0.4914, don’t reject \( H_0 \). There isn’t statistically significant evidence that the treatment means differ.

P5. The boxplots of Treatments 4 and 5 do not overlap; they appear to have widely differing means.\( (18.4, 30.2) \) does not include 0, so reject \( H_0 \) that the mean growth would have been the same if all plants could have been given both treatments.

P6. Outliers for Treatment 4: 1 and 4; outlier for Treatment 5: 5; Treatment 5 is skewed right, whereas Treatment 4 is fairly symmetric. Plants given Treatment 5 were slightly taller and more variable in height to begin with, but there isn’t enough difference to question the conclusion in P5.

**Section 12.2**

P7. Conditions are met. With \( z = -0.7303 \), the one-sided \( P \)-value is 0.2326. Do not reject the \( H_0 \) that the proportion of Florida households in which the primary language is not English is 0.25 in favor of the alternative that it is less.

P8. Conditions are met. You are 95% confident that the proportion of all Florida households residing in an apartment or other attached structure is in the interval \( 0.30 \pm 0.14 \), or (0.16, 0.44).

P9. Conditions are nearly met. Any population proportion in the interval \( 0.225 \pm 0.129 \), or (0.096, 0.354), could produce the sample proportion as a reasonably likely outcome.

P10. For those who own their home with a mortgage and those who own their home outright, there is no clear association with building type. Renters appear more likely to live in an apartment or other attached structure.
P11. Conditions are not quite met, as one of the four cells has an expected frequency less than 5. With $\chi^2 = 0.2564$ or $z = 0.5064$, the $P$-value = 0.6126. Don’t reject $H_0$ that the proportion of households with older residents is the same for both language categories.

P12. Conditions are not met, as three of the cells have expected frequencies below 5. With $\chi^2 = 6.8899$ and $P$-value = 0.0319 (perhaps a little larger because of failed conditions), reject the $H_0$ of no association. There is a statistically significant association between the type of building in which the household resides and the primary language spoken.

P13. a. The $BLD$ values were randomly paired with the values of $HHL$. The frequency of each type of pair was recorded, expected frequencies calculated, and $\chi^2$ statistic computed. This was repeated 200 times.

b. Yes. The observed test statistic 6.8899 was met or exceeded 7 times out of 200 runs, making the approximate $P$-value 0.035.

P14. Conditions are met. You are 90% confident that the mean number of people per household in Florida is in the interval $2.175 \pm 0.301$, or $(1.874, 2.476)$.

P15. a. There is wide spread but the distribution is symmetric with no outliers, so with a sample size of 24, the standard method based on the $t$-distribution should work without a transformation.

b. With $t = -0.8105$ and $P$-value = 0.213, do not reject $H_0$. You do not have statistically significant evidence that the mean mortgage payment in Florida is less than $840$.

P16. a. The distribution is highly skewed with three outliers. A log transformation works well.

b. Conditions are met reasonably well for the transformed data. You are 95% confident that the mean of the natural log of the cost of hazard insurance, for all households that pay it in Florida, is in the interval $6.510 \pm 0.334$, or $(6.176, 6.844)$.

P17. Conditions are met, except that $rooms$ for apartments is a bit skewed. You are 95% confident that the difference in the mean number of rooms in detached houses and the mean number of rooms in apartments is in the interval $1.82 \pm 0.97$, or about $(0.85, 2.79)$.

P18. If one house has one more room than another house, you would expect an increase in the electricity bill of about $17.10$ per month.

P19. Yes. Conditions are met reasonably well. With $t = 3.120$, the (one-sided) $P$-value is about 0.0017. A one-sided test can be justified because additional rooms should raise the electric bill.

P20. The scatterplot shows a statistically significant linear trend ($P$-value = 0.00076), but the correlation is not high. This is strong evidence of a fairly weak association.

P21. a. Both scatterplots indicate a weak positive trend, but the number of rooms has slightly higher correlation with the number of vehicles.

b. Conditions are met for both tests.

People: With $t = 1.97$ and $P$-value = 0.0564, the observed slope is not quite significantly different from 0 at the 5% level.

Rooms: With $t = 2.85$ and $P$-value = 0.069, the observed slope is significantly different from 0 at the 5% level. Number of rooms is the better predictor. This could be because wealth (a lurking variable) is more closely related to number of rooms than number of people in a household.

P22. a. $INSP = -618.41 + 289.662 \cdot RMS$. For every additional room in the house, the annual insurance payment tends to increase by about $290$.

b. The main problem is the lack of homogeneity.

c. A log transformation on $INSP$ fixes this quite nicely. $ln(INSP) = 4.68 + 0.322 \cdot RMS$. You are 95% confident that the slope of the linear relationship between the natural log of a household’s annual hazard insurance payment and the number of rooms in the household for all Florida households is in the interval $0.322 \pm 0.1933$, or $(0.129, 0.515)$.

Section 12.3

P23. a. A statistically significant positive linear relationship ($t = 4.56; df = 28; P$-value $< 0.0001$). On the average, there is an $22,700$ increase in payroll for each increase of 1,000 in attendance.

b. The American League mirrors essentially the same pattern as both leagues together, with a statistically significant ($t = 3.313, df = 12,$
P-value = 0.0062) increasing trend of about $26,700 in increased payroll for each increase of 1,000 people in attendance.

c. The scatterplot indicates serious curvature, which isn’t straightened by our usual transformations. A more complex model is needed.

P24. For all three groups of teams, there appears to be little association between payroll and batting average. P-values are
a. 0.3378  b. 0.4760  c. 0.6208

P25. a. A statistically significant positive linear relationship; if the two teams with unusually high batting averages (Colorado and Seattle) are removed, the estimate of the slope increases from 3.72 to 4.5 tenths of a percentage point average gain in percent wins per 1-point increase in team batting average and the P-value remains about the same.

b. The average increase in percent wins (in tenths) is 7.2 per 1-point increase in batting average. However, if Seattle is dropped from the analysis, the estimate of the slope, 5.3, is not quite statistically significant at P-value = 0.071.

c. With Colorado, there is no significant linear relationship. Without Colorado, the slope jumps to an estimated 5.8 increase in percent wins (in tenths) for each 1-point increase in batting average and the P-value drops to 0.006.

P26. a. No; t = -0.5263 and P-value = 0.6. The difference in mean attendance is not statistically significant.

b. No; t = 1.447 and P-value = 0.16. The difference in the mean batting average is not statistically significant.

P27. No; z = 0.4503 and P-value = 0.6525. A difference this small in winning percentages can reasonably be attributed to chance.

P28. Use a chi-square test of homogeneity. The outcomes are not strictly independent; however, of the 162 games played, only a small number are played among each other, so the dependence is not too serious.

a. No; \( \chi^2 = 0.351 \) with \( df = 2 \), P-value = 0.84. The observed differences could very well be produced by chance alone.

b. Yes; \( \chi^2 = 9.203 \) with \( df = 2 \), P-value = 0.01. The differences in the percentages of wins cannot reasonably be attributed to chance alone.

**Section 12.4**

P29. Youth Enterprises; the variance of the difference is minimized when sample sizes are equal.

P30. a. Min = 22, Q1 = 37, Median = 53, Q3 = 56, Max = 69

b. Boxplot shows that there are no outliers. Stemplot shows the distribution is bimodal, with one peak in the low 30’s and a second, higher peak in the upper 50’s. Both show skewness.

P31. 72%; 56%

P32.

<table>
<thead>
<tr>
<th>Age</th>
<th>Terminated?</th>
<th>Under 50</th>
<th>Yes</th>
<th>Total</th>
<th>Percentage Terminated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under 50</td>
<td>13</td>
<td>9</td>
<td>22</td>
<td>40.91</td>
<td></td>
</tr>
<tr>
<td>50 or Older</td>
<td>9</td>
<td>19</td>
<td>28</td>
<td>67.86</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>22</td>
<td>28</td>
<td>50</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

P33. With \( z = 1.9055 \) the one-sided P-value = 0.0284. Reject the \( H_0 \) that the difference in the proportion of workers laid off who were under 50 and the proportion of workers laid off who were 50 or older is no larger than you would expect if Westvaco were picking 28 people at random for layoff.

P34. Using 50 as a cutoff leads to much stronger evidence in support of a claim of discrimination.

P35. a. Neither test so far meets the criterion established by the Supreme Court, although the second test (in P33) comes very close with a one-sided P-value of 0.0284.

b. I

c. A Type I error, if the null hypothesis is true
d. A Type II error would be deciding that a company does not discriminate on the basis of age when, in fact, it does.

P36. The distributions are somewhat skewed toward the smaller ages. The average age of those terminated is higher than the average age of those retained. The spread of the ages of those terminated is slightly greater than for those retained.
P37. Here are the five-number summaries needed to make the boxplots.

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terminated</td>
<td>22</td>
<td>38.5</td>
<td>54.5</td>
<td>59</td>
<td>69</td>
</tr>
<tr>
<td>Retained</td>
<td>25</td>
<td>37</td>
<td>48</td>
<td>55</td>
<td>61</td>
</tr>
</tbody>
</table>

P38. With $t = 1.066$ and $P$-value = 0.146, there is insufficient evidence to reject $H_0$. If 28 workers were selected totally at random to be laid off, then it is reasonably likely to get a difference of 3.68 years or more in the average age of employees laid off minus the average age of the employees retained.

P39. a. The hypotheses are the same. In both cases the null hypothesis is that the actual result can reasonably be attributed to chance, and the alternative hypothesis is that it cannot.

b. For a $t$-test, you must check that both samples look like they could reasonably have come from a normally distributed population. This is not necessary with a randomization test.

c. For a $t$-test, if the conditions are met, the sampling distribution should have a $t$-distribution. The sampling distribution for a randomization test is difficult to predict, but often looks somewhat normal.

d. If the conditions for a $t$-test were met, the two tests should reach the same conclusion. The $P$-values would be very close. If the conditions have not been met, the randomization test is more reliable.
index

A
absolute value, spread and, 65
Addition Rule, 319–320
  for Disjoint Events, 316–318
for random variables, 372–375
adjacent values, 62
Adrain, Robert, 123
age-neutral simulation, 13–15
d'Alembert, Jean, 291–292
alternate hypothesis
  for difference of two means, 625, 647
  for difference of two proportions, 530, 532
  for experiments, 544
  forms of, 498
  for means, 589
  one-sided, 594–595
  for one-sided tests, 507, 508
  for surveys, 530, 532
See also hypotheses; null hypothesis
American Community Survey (ACS)
case study, 803–808
anecdotal evidence, 249–250
Anscombe data, 166–167
approximate (simulated) sampling
distribution, 410–411, 417, 428–429
area, standard deviation and, 32
associated variables, 711, 722
association. See strength of association
average. See mean
averages, line of. See least squares regression line
averages, point of, 125

B
back-to-back graphs, 48, 55
balance point. See mean
balanced design, 266
bar graphs, 50–51
  histograms compared to, 50
  stacked or segmented, 693, 713–714
Barry, Dave, 226, 348–349
bars
  of bar graphs, 50
  of histograms, 44–46
baseball salaries case study, 810–816
Bernoulli, Jacob, 293
bias, 220–227
  convenience sample, 224
  defined, 222
  frame of sample and, 224–225
incorrect response or measurement, 226–227
  judgment sample, 224
  of method vs. sample, 222–223
  nonresponse bias, 225
point estimators and, 415, 416
questionnaire bias, 225–226
response bias, 225–227
sample selection bias, 223–225
size bias, 223
  voluntary response bias, 224, 246
See also confounding; random sampling
bimodal distributions, 35–36, 37
binomial distributions, 382–390
calculator technique, 385, 389
defined, 385
expected value of, 387–388
independence and, 383, 385, 386
shape of, 388
standard deviation of, 388–389
bins, 44
bivariate categorical data, 805–807
bivariate quantitative data, 107–110
  heteroscedasticity and, 774–775, 778
  inference with, 807–808
  outliers and, 107, 108, 162
  regression and, 739
blind experiment, 251
blocking, 269, 273
blocks, 267, 268–270, 273, 274
box-and-whiskers plots. See boxplots
boxplots, 61, 62–63
  calculator technique, 63
C
Calculator Notes references
  binomial distributions, 385–386, 389
central intervals for normal distributions, 92
chi-square probability density function, 678
chi-square test of homogeneity, 699
conditions for significance test, 757
confidence interval for a mean, 571
confidence intervals, 476, 477
correlation, 141
critical values, 497
cumulative relative frequency plots, 79
deviation, random selection of, 740
dot plots, 4
frequency tables, 68, 290
general distributions, 394, 396
graphing by specifying mean and standard deviation, 33
least squares regression line, 124, 126, 131, 141
normal distributions, 33, 92
numbers of combinations, 20
P-values, 496, 585, 680, 718
pairs of random integers, generating, 305
predicted values, 122
probability distribution, 367
random digit string, generating, 301, 449
random number distribution, 30
random number generation, 234, 305, 563
residuals, 122, 131
sampling distributions, 412
scatterplot, 107, 121, 131
simulation, 14, 306
spinning coins, 293
sum of the squared errors (SSE), 124
t-distributions, 584
t*-values, 566
t test statistic, 718
transformations, 192, 605
two-sample t-interval, 622, 626
unit changing, effects of, 76
z-scores, 84, 85, 89

calculators
  histograms and, 45, 46
  least squares lines on, 127–128
  percentiles on, 85
  standard deviation and, 65
  standard normal distribution, 84
  summary statistics and, 66
See also Calculator Notes references; software, statistical
capture rate, 477–479
case studies
  American Community Survey (ACS), 803–808
  baseball salaries, 810–816
flower production, 800–803
Westvaco case, 3–8, 11–16, 817–822
cases
of bivariate data, 107
defined, 4
categorical data, 692–693, 805–807
categorical variables
defined, 44
graphs of. See bar graphs
inference involving. See chi-square; proportion
strength of association of, 722–723
cause and effect
correlation not implying, 147
experiments establishing.

central intervals, 90–92
census, 219, 358
Celsius, conversion to, 76
Central Limit Theorem, 430–431
center
defined, 27
of normal distributions, 31
recentering. See recentering
skewed distributions and, 34
center, measures of
computation of, 56–58
mean. See mean; mean of a population; sample mean
median. See median
midrange, 73, 421
central intervals, 90–92
Central Limit Theorem, 430–431
chance. See random sampling
chi-square
calculator technique, 678
critical values, table of, 827
degrees of freedom and, 679–680, 695–696, 703–705
distribution of, 677–678, 684
expected frequency of, 684, 694–695, 715
goodness-of-fit test, 674, 680–683, 684–686
history of, 674
homogeneity test. See homogeneity, chi-square test of
independence. See independence, chi-square test of
sample size and, 678, 722–723
stacked bar graph, 693, 713–714
symbol ($\chi^2$), 675
table of values for, 679–680, 829
test statistic, 674–675, 681
two-way table and, 692–693, 715
z-test compared to, 685, 702–705
See also significance tests
clinical trials, 247, 251, 542–543, 545–546
See also experiments
cluster samples, 237–238, 240
clusters, 36–37, 107
coefficient of correlation. See correlation
coefficient of determination, 148–151
column chart, 714
column marginal frequency, 713
common ratio, 401
comparative studies. See experiments; observational studies
comparing, as characteristic of experiment design, 252
comparisons group, 250–251
complement, 288
completely randomized design, 264, 265–266, 274, 643
conditional distribution of $y$ given $x$, 741
conditional probability, 325–333
definition of, 330
inference and, 322–333
medical tests and, 330–331
Multiplication Rule and, 327–329, 330
from the sample space, 326
symbol for, 326
conditional relative frequencies, 693, 713
conditions assigned. See factors; treatments
conditions of significance tests

calculator technique, 757
checking, 500
for difference of two means, 625, 647
for difference of two proportions, 531–532, 544
fixed-level test for mean, 589
goodness-of-fit, 681, 682
homogeneity, 697, 699
independence, 716, 718
for means, 589, 625
for one-sided tests, 508
for proportions, 498, 500, 509
for slope, 757, 758, 759–760, 765
t-test, 820–821
See also significance tests
confidence intervals, 470–473, 482–483
95% confidence interval, 473, 572
calculator technique, 476, 477
capture rate, 571–572, 604
chart for reasonably likely events, 470–472
conditions required for, 476
correct language use of, 479
for difference in two proportions, 518–523, 541–543
for difference of two means, 618–623, 629
for difference of two means from paired comparisons, 645–646
error sources, 430
estimation of standard deviation and, 563–568
for experiments, 541–543
formulas for, 473–477, 519–520
margin of error. See margin of error
for mean of population, 562–573
for mean of samples. See sample means, confidence intervals for
for observational studies, 547–548
outliers and, 604–609
for proportions, 429–430, 475–477
skewness and, 602–609
for slope, 762–763, 765
for surveys, 518–523
See also significance tests
confidence interval, $t$-values and, 566
confounding
in comparative studies, 245–247, 248–249
See also bias; random sampling
constant strength, 107
constant values, 76
case group, 250–251
correlation sample, defined, 224
correlation, ecological, 815–816
correlation ($r$), 140–151
and appropriateness of linear model, 144–145
cause and effect not implied by, 147
and coefficient of variation ($r^2$), 148–151
estimating, 140–141
formula for, 142–144
interpretation of, 148–151
negative, 140
positive, 140
slope, relationship with, 146, 152
steps for determining, 107–110
strength of, 140
as term, 140
transformations vs. use of, 779
units absent from, 143
See also strength of association
count. See frequency, histograms and Counting
Fundamental Principle of, 294–295
critical values
calculator technique, 497
chi-square distribution, table of, 679–680, 829
of confidence interval for a mean, 475–476
defined, 496
and level of significance, 496–497
t-distribution, table of, 828
test statistic and, 496–497
See also t*-values
cumulative relative frequency plots, 78–79
cumulative probabilities, 385–386, 396
cumulative relative frequency plots, 78–79

data
bivariate. See bivariate categorical data; bivariate quantitative data
categorical, 692–693, 805–807
eLLiptical, 154
independence and. See independence, chi-square test of independent events with, 343–345
inference requirements and, 651, 652, 664
probability distributions from, 358–360, 361
production of, 640
quantitative (measurement), 807–808
production of, 640
probability distributions from, 358–360, 361
production of, 640
quantitative (measurement), 807–808
raw, 27
summaries. See summary statistics
symmetry and, 292
time-series, 765
univariate, 805–807
See also paired comparisons
decay, exponential, 179–180
degrees of freedom
chi-square and, 679–680, 695–696, 703–705
difference of two means and, 621
slope estimation and, 756, 764–765
t-values and, 566, 584, 621
dependent events
defined, 339
matched pairs design and, 644
time as, 765
See also independent events
dependent variables, 711
design. See experiments and experiment design
determination, coefficient of, 148–151
deviations, 64–65
See also standard deviation
diagnostics
outliers and, 162–167, 172
residual plots, 168, 172
vs. summaries, 162, 167, 172
diagrams, tree, 294–295, 327–328
difference of two means
confidence intervals and, 618–623, 629, 645–646
increasing power in, 629–630
significance testing and, 624–628, 629–630, 646–647, 820–821
difference of two proportions
confidence interval for, 518–523, 541–543
sampling distribution of, 527–529, 539–541
significance test for, 526–535, 544–546, 819
difference, sampling distribution of, 445, 527–529
discrete distributions, chi-square, 684
disjoint events (mutually exclusive events), 315–316
Addition Rule and, 316–318
defined, 291
property of, 317
distributions
centers of. See center, measures of chi-square. See chi-square conditional, of y given x, 741
defined, 4
modes of, 35–36
probability. See probability distributions
recentering. See recentering
rescaling. See rescaling
sampling. See sampling distributions
shapes of. See bimodal distributions;
normal distributions; skewed
distributions; uniform (rectangular) distributions
summaries of, 37
See also graphs and graphing
dot plots, 44
calculator technique, 4
defined, 4
mean, estimation of, 57
rounding rules and, 44
software for statistics and, 44
usefulness of, 44
double-blind experiment, 251

e
ecological correlations, 815–816
ecological fallacy, 816
ecnomics, sampling and, 219, 237, 480–481
educational research, units and, 252
eLLiptical data, 154
errors
degrees of freedom and, 703–705
hypotheses and, 501–506, 593–594
measurement error problems, 592
prediction, 120
sum of squared, 123–125, 128
See also margin of error; standard error; standard error of the mean
estimation
of mean, 57
of median, 34
of parameter, 220
point estimators, 415–417
of slope, 756, 764–765
of standard deviation, 32, 563–568
units of measurement and, 50
ethics, observational studies and, 547
even numbers, and quartiles, finding, 59
events
defined, 288
dependent. See dependent events
disjoint. See disjoint events
independent. See independent events
rare, 412, 468, 593, 761
See also reasonably likely events;
sample space
exact sampling distributions, 413–414
expected frequency, of chi-square tests, 684, 694–695, 715
expected value
of binomial distributions, 387–388
defined, 365
in everyday situations, 369–370
formula for, 366
of geometric distributions, 397–398
linear transformations on, 371–375
standard deviation and, 364–367
symbol for, 365
experiments and experiment design
balanced, 266
blind, 251
cause inferred through, 243–245, 256, 279
characteristics of, 252–256
clinical trials, 247, 251
completely randomized design, 264,
265–266, 274, 643
certainty intervals of, 541–543
confounding in, 245–247, 248–249
control groups, 250–251
double-blind, 251
evidence, strength of, 617–618
flower production case study,
800–803
lurking variables in, 243–244
matched pairs design, 264–265, 643–644, 645
observational studies contrasted to, 247, 279
placebo effect, 250–251
protocol, 270
random assignment, 539–541
randomized block design, 268–270, 274
randomized comparative experiment, 251
randomized paired comparison design, 264–265, 266–268
randomizing. See randomizing
repeated measures design, 265, 644–645
replication, 252, 253
response in, 244
significance tests for, 544–546
surveys contrasted to, 278–279, 539
as term, 245
treatments. See treatments
units. See units, experimental variability, management of, 262–263, 269, 270–274
See also observational studies; random sampling; surveys
experts. See judgment sample explanation of pattern, 107, 109
explanatory (predictor) variable, 120
exploration of data, defined, 3
exponential functions
growth and decay, 179–186
and log transformations, 180–187, 193
regression lines and, 774, 781
extrapolation, 120, 129

F
factors, 248
Fahrenheit, conversion to Celsius, 76
fallacy, ecological, 816
false negatives, 331
false positives, 331
first quartile. See lower quartile
Fisher, R. A., 15, 269
Fisher's exact test, 819
five-number summary, 60–61
calculator technique, 63
fixed-level testing, 588–593
flower production case study, 800–803
frame, 224–225
frequency tables
calculator technique, 68, 290
chi-squares and, 692–693, 712–713
summarizes from, 67–68
See also relative frequency
histograms; relative frequency plots
Fundamental Principle of Counting, 294–295

G
Galton, Francis, 152
gaps, 36–37
Gauss, Carl Friedrich, 123
generalization of pattern, 107, 108
genes, 490
geometric distributions, 393–399
calculator technique, 394
defined, 396
expected value of, 397–398
formula for, 395–396
memoryless property of, 404
shape of, 398
standard deviation of, 398
as waiting-time problem, 393
geometric sequence, 401
glossary, 829–838
false positives, 331
false negatives, 331
fallacy, ecological, 816
Fahrenheit, conversion to Celsius, 76
factors, 248
Fundamental Principle of Counting, 294–295

H
heteroscedasticity, 107, 774–775, 778
histograms, 44–47
back-to-back, 55
bar graphs compared to, 50
bars of, 44–46
bins of, 44
calculator technique, 46
choice of, summarized, 51–52
column chart, 714
cumulative frequency plots, 78–79
dot plots. See dot plots
histograms. See histograms
independence tests and, 712–714
questions raised by, 50
scatterplots. See scatterplots
stemplots. See stemplots
growth, exponential, 179, 182–186
hypotheses
alternate. See alternate hypothesis
errors and, 501–506, 593–594
for goodness-of-fit test, 681, 682–683
for homogeneity test, 697, 698–699
for independence test, 716, 718
for means, 585–587, 589, 625
null. See null hypothesis
for proportions, 498, 509
for slope, 758, 760
for t-tests, 589
See also significance tests
hypothesis tests. See significance tests
icorrect response bias, 226–227
independence, chi-square test of, 711–724
calculator technique, 718
expected frequencies in, 715
graphical display of data, 712–714
homogeneity test vs., 720–722
procedure for, 716–719
sample size and, 722–723
strength of association and, 722–723
independent, as term, 345
independent events, 339–345
binomial distributions and, 383, 385, 386
calculator technique, 68, 290
chi-squares and. See chi-square
data and, 343–345
definition of, 339–340, 345
definition of, 383, 385, 386
definition of, 396
Multiplication Rule for, 341–343
random variables, 372–373
randomization and, 249
data requirements for, 651, 652, 664
defined, 3
observational studies and, 547, 812
with quantitative data, 807–808
randomization and, 249
significance testing and. See
significance tests
simulation and, 12–16
See also cause and effect;
experiments; observational studies; random sampling
inflection points, 31–32
informed consent, 542
interpolation, 120, 129
interquartile range (IQR), 59–60, 61–62
IQR. See interquartile range
J
joint frequency, 693, 713
judgment sample, 224
L

Law of Large Numbers, 293
least squares regression line, 123
calculator technique, 124, 126, 131, 141
computer output for, 127–128
equation, 125
finding, 123–124, 125–126
as line of means, 152–153
linear models and, 739–741, 772–773
point of averages of, 125
properties of, 125
slope of, 125, 126, 128
sum of squared errors (SSE), 123–125, 128
as term, 152
transformations and, 189, 774–779, 781
true regression line equation, 739
y-intercept of, 125, 126, 128, 739
See also slope
leaves, 48
Legendre, Adrien-Marie, 123
Leonardo da Vinci, 141
level of confidence, 566
level of significance, 496–497, 500, 588
levels of factors, 248
line of averages. See least squares
regression line
line of means. See least squares
regression line
line of symmetry, 31, 32
linear equations
finding, from two points, 118–119
slope-intercept form, 116–119
transformations to find. See transformations
linear models
checking fit of, 771–773
correlation and appropriateness of, 144–145
selection of, 188
true regression line and, 739–741, 772–773
linear transformations, 75–76, 371–375
linearity, bivariate data and, 107, 108 lines
least squares. See least squares
regression line
for prediction, 120–122
regression. See least squares
regression line
as summaries, generally, 116–119, 129–130
Lister, Joseph, 707
log-log transformations, 187–189, 774–779

logarithmic transformations, 193
exponential functions and, 180–187, 193
power transformations and, 13, 187–189, 774–779
skewed distributions and, 605
See also exponential functions
logistic functions, 779
lower quartile
defined, 34
in five-number summary, 61
interquartile range and, 59
lurking variables, 107, 109, 147, 243–244

M
margin of error
confidence interval of a mean and, 572–573
defined, 476–477
sample size and, 479–481, 572
marginal frequencies, 693
marginal relative frequency, 693
matched pairs design, 264–265, 643–644, 645
maximum
adjustment of sample maximum, 426
in five-number summary, 61
range and, 61
skewed distributions and, 34
Mayo, Charles, 246, 247
mean (\(\mu\))
calculator technique, 66
computation of, 56
confidence intervals for. See under mean of a population; sample mean
defined, 32, 56
deviations from, 64–65
estimation of, 57
of frequency tables, 67–68
midrange and, 73
of a population. See mean of a population of a probability distribution. See expected value
regression toward, 152–153
of a sample. See sample mean of a sampling distribution, 430, 435, 459
as sensitive to outliers, 79
significance tests for. See sample means, significance tests and; t-tests
standard error of. See standard error of the mean
as summary statistic, choice of, 74
trimmed, 73
mean of a population (\(\mu\))
capture rate for, 571–572
certainty intervals for, 562–573
margin of error for, 572–573
point estimators, 415–417
significance testing of, 580–596. See also t-tests
symbol for, 430
test statistic for, 581–582, 589, 625
means, line of. See least squares regression line
measurement
inaccurate, bias and, 227
units of. See units of measurement
measurement data, 807–808
measurement error problems, 592
measures of center. See center, measures of median
computation of, 57–58
defined, 57
estimation of, 34
as resistant to outliers, 79
of skewed distribution, 34
spread around (interquartile range), 59–60, 61–62
medical experiments. See experiments medical tests, 330–331
midrange, 73, 421
minimum
in five-number summary, 61
range and, 61
skewed distributions and, 34
mode
defined, 31
of normal distributions, 31
two or more, 35–36
models. See linear models; probability models
modified boxplot, 62–63
Multiplication Rule, 327–329, 330
for independent events, 341–343
mutually exclusive outcomes. See disjoint events

N
negative trend, 107
nonresponse bias, 225
normal distributions, 30–33
binomial distributions and, 388
calculator technique, 33
center of, 31
central intervals for, 90–92
common sources of, 33
certainty interval of a mean and, 568, 569
defined, 31
difference of two means from paired observations and, 645, 647
null hypothesis, 492
unknown value problems and, 84–85
unknown percentage problems and,
table of probabilities, 826–827
symmetry line of, 31, 32
summary of, 37, 74, 92
standard.
spread of, 31
significance test for a mean and, 589
shape of, 31, 83
surveys contrasted to, 279
significance test for, 548–549
levels of factors, 248
inference and, 248
confounding in, 245–247, 248–249
defined, 246
ethics and, 547
experiments contrasted to, 247, 279
factors in, 248
inference and, 547, 812
levels of factors, 248
significance test for, 548–549
surveys contrasted to, 279

P
P-values
calculator technique, 496, 585, 680
computing, 495–496
defined, 494
for difference of two means, 625, 647, 650
for difference of two proportions, 531, 532, 544–545
for means, 585–587, 589, 625
defined, 36, 61
influence of, 77, 162–167, 172
and interquartile range, value of,
61–62
scatterplots and, 162–167
See also hypotheses

O
observational studies
baseball salaries case study, 810–816
confidence interval for, 547–548
confounding in, 245–247, 248–249
defined, 246
ethics and, 547
experiments contrasted to, 247, 279
factors in, 248
inference and, 547, 812
levels of factors, 248
significance test for, 548–549
surveys contrasted to, 279
units and, 252–254
See also experiments; surveys
observed significance level. See
P-values
odd numbers, and quartiles, finding, 59
one-sided (one-tailed) tests of
significance, 507–508, 594–595
opinion polls
methods of, 468
questions to ask about, 482
See also surveys
“or,” as term, 315
outcomes
disjoint. See disjoint events
equally likely, 288
Fundamental Principle of Counting,
294–295
list of. See sample space
See also events
outliers, 36–37
bivariate data and, 107, 108, 162
boxplots and, 62–63
certainty intervals of a mean and,
604–609
defined, 36, 61
and interquartile range, value of,
61–62
scatterplots and, 162–167

P
power of test and, 503–506, 594,
629–630
for one-sided tests, 507, 508
one-tailed, 594–595
for experiments, 544–545
for homogeneity test, 696, 697–698
for independence test, 716–717, 718
interpretation of, 494–495, 496
for means, 585–587, 589–590,
594–595
for one-sided tests, 507, 508
one-tailed, 594–595
simulation of, 814
slope and, 761
for surveys, 531, 532
confidence intervals of a mean and,
644, 645–646
definition interval and, 644,
645–646
definitions of experiment and, 540–545
meaningful comparisons, 650–651
significance tests and, 646–647
transforming non-normal data,
648–649
parameter, 220
Pearson, Karl, 674
percentiles, 78–79
permutation test, 821
placebo, 250–251
placebo effect, 250
plausible percentages, 471–472
plausible values, 562–563
point estimators, 415–417
point of averages, 125
pooled estimate, 530, 629–630
population
census of, 219, 358
defined, 219
frame contrasted to, 224–225
mean of (μ). See mean of a
population parameter of, 220
sampling of. See random sampling
size of (N). See population size
standard deviation of. See
population standard deviation (σ)
known, 358
population pyramid, 55
population size (N)
defined, 219
replacement and, 326, 433
representativeness of sample and,
433–434, 479–481
sampling distribution of the mean
and, 434
symbol for, 430
population standard deviation (σ)
defined, 412
estimating, 563–568
reasonably likely events and, 412
symbol for, 430
See also standard deviation; standard
error
positive trend, 107
potentially influential outliers, 163–167
power functions, equation of, 188
power of tests, 503–506, 594, 629–630
power transformations, 190–192, 193
logarithms and, 187–189, 774–779
regression lines and, 774, 781
predicted (response) variable, 120
predicted value, 120, 130
prediction
calculator technique, 122
error of, 120
extrapolation, 120, 129
interpolation, 120, 129
lines used for, 120–122
sample size vs. population size and, 326, 433
sampling distribution properties and, 433, 444–445
replication, 252, 253
rescaling, 75–76, 79
of probability distributions, 371–375
standardizing with, 83, 85–87
See also transformations
residual plots, 167–172
calculator technique, 169
construction of, 168–169
defined, 168
as diagnostic plot, 168, 172
interpretation of, 170–171
time and, 184, 193
types of, 171–172
residuals
calculator technique, 122
defined, 120
finding, 120–122
homogeneity of, in power model, 188
regression slope and, 77
as sum of squared errors, 123–125, 128
See also residual plots
resistance to outliers, 77
response, 244
response bias, 225–227
response variable, 120
robustness, 602, 607
row marginal frequency, 713

S
sample mean (\(\bar{x}\))
comparing, usefulness of, 650–651
confidence intervals. See sample means, confidence intervals and importance of, 426
point estimators, 415–417
finding probabilities of, 431–433
reasonably likely outcomes for, 412
sample proportion as form of, 450
sampling distribution of. See sampling distribution of the sample mean
significance tests for. See sample means, significance tests and symbol for, 430
use of, in text, 426
sample means, confidence intervals and, 573
capture rate and, 571–572
collection of, 568–571
for difference between two means, 618–623, 629
differences, paired comparisons, 645–646
estimation of standard deviation and, 565–568
sample size, 568, 569, 572
skewness and outliers, effect of, 602–609
sample means, significance tests and, 580–596
components of, 589–590
for difference between two means, 624–628, 629–630
differences, based on paired samples, 546–547
fixed-level testing, 588–593
one-tailed tests, 594–595
P-value and, 585–587, 594–595
rejection of null hypothesis, meaning of, 593–594
sample proportion (\(\hat{p}\))
assigning variables, 521
Law of Large Numbers and, 293
mean and, 450
pooled estimate, 530, 629–630
probabilities of, finding, 452–453
reasonably likely, 468–469
sampling distribution of, 448–453
standard deviation and, 450–451
symbol for, 448
testing. See significance tests
sample selection bias, 223–225
sample size
binomial distributions and, 386
Central Limit Theorem and, 431
chi-square and, 678, 722–723
confidence interval of a mean and, 568, 569, 572
difference of two means and, 621, 647
economics and, 219, 480–481
estimation formula, 480–481
Fisher's exact test, 819
listing samples, generation by, 410, 417
mean of, 433, 434, 437
marginal error and, 417
margin of error and, 479–481
power and Type II errors and, 530–506, 594, 629
replacement and, 326, 433
repli...
Index 891
of frequency table, 67–68
of geometric distributions, 398
model checking and, 772–773
of populations. See population
standard deviation (σ)
of probability distributions, 364–367, 371–375
of sample proportion, 450–451
of sampling distributions. See
standard error (SE)
as sensitive to outliers, 79
as summary statistic, choice of, 74
symbols for, 430
variance and. See variance standard error of the mean
defined, 430
sample size and, 431
symbol for (σx), 430
standard error (SE)
defined, 412
of difference of two proportions, 519–520
of estimated slope, 741–748
of the number of successes, 447
point estimators and, 415
of the sample proportion, 450
of sampling distributions, 435, 459
of the sum of a sample, 435
standard normal distributions, 83–85
standard normal probabilities, table of,
826–827
standard units. See z-score
standardizing, 83–85
defined, 83
solving unknown percentage and
value problem, 87–90
standardizing, 83, 85–87
table of probabilities, 826–827
unstandardizing, 86
See also z-score
standard normal probabilities, table of,
824–825
standard sampling, 435
defined, 56
of probability distributions, 531, 532, 544
defined, 565
degrees of freedom and, 566, 584, 621
outliers and, 604
P-values and, 593–594
significance testing and, 624–628,
760–761
t-distributions, 582–584, 828
t-tables, 565–566, 828
z-values compared to, 583
tables
chi-square values, 679–680, 829
frequency. See frequency tables
probability distribution, 365–366,
375, 824–825
random digits, 301–309, 359–360,
830
random number, 232
reading vertically, 4
summary, 6
t-tables, 565–566, 828
two-way, 692–693, 715
z-table, 84–85, 826–827
tail probability, 586
test statistic (t)
chi-square tests, 674–675, 681
critical values and, 496–497
defined, 493–494
for difference of two means, 625,
647, 650
for difference of two proportions,
531, 532, 544
error types and, 501–502, 593–594
of experiments, 544
goodness-of-fit test, 681, 683
homogeneity test, 695–696, 697, 699
for independence test, 716, 718
for mean of a population, 581–582,
589, 625
normal distribution of, 581
for one-sided tests, 508
paired differences, 647

of frequency table, 67–68
of geometric distributions, 398
model checking and, 772–773
of populations. See population
standard deviation (σ)
of probability distributions, 364–367, 371–375
of sample proportion, 450–451
of sampling distributions. See
standard error (SE)
as sensitive to outliers, 79
as summary statistic, choice of, 74
symbols for, 430
variance and. See variance standard error of the mean
defined, 430
sample size and, 431
symbol for (σx), 430
standard error (SE)
defined, 412
of difference of two proportions, 519–520
of estimated slope, 741–748
of the number of successes, 447
point estimators and, 415
of the sample proportion, 450
of sampling distributions, 435, 459
of the sum of a sample, 435
standard normal distributions, 83–85
defined, 83
solving unknown percentage and
value problem, 87–90
standardizing, 83, 85–87
table of probabilities, 826–827
unstandardizing, 86
See also z-score
standard normal probabilities, table of,
824–825
standard units. See z-score
standardizing, 83–85
statistic, as estimate from sample, 220
statistical significance, 490, 492, 501
stem-and-leaf plot. See stemplots
stemplots, 47–50
back-to-back, 48
rational rules and, 49
software for statistics and, 49
split stems, 48
truncation of values, 49
usefulness of, 49
stems, 48
strata, defined, 235
stratification, defined, 235
stratified random sample, 234–237,
240
strength of association
correlation and. See correlation
defined, 722

independence tests and, 722–723
scatterplots and, 107, 108
strength of evidence, 617–618
Strutt, John William, 36–37
sum (Σ), as symbol, 56
sum of squared errors (SSE), 123–125,
128
summary statistics
bias and, 415, 416
of bivariate data, 107–109, 116–119
calculator technique, 66
choice of, generally, 4–5, 28, 74–75
defined, 12
diagnostics vs., 162
distribution of. See sampling
distributions
of distributions, 37
five-number, 60–61, 63
outliers, influence of, 77
point estimators, 415–417
precise, 415
proportion. See sample proportion
recentering of, 75–76, 79
rescaling of, 75–76, 79
See also center, measures of;
diagnostics; spread
summary tables, 6
surveys
American Community Survey
(ACS) case study, 803–808
bias and. See bias
causal inferences and, 278, 279
eperiments contrasted to, 278–279,
539
parameters and, 220
random sampling for. See random
centering of, 75–76, 79
random sampling for. See random
sampling
samples used in, reasons for,
219–220
significance testing for, 529–534
See also experiments; observational
studies
symbols, glossary of, 831–832
symmetry
defined, 33
line of, 31, 32
of normal distribution, 31
probability and, 289, 290, 292
systematic sample with random start,
238–239, 240
T
t-distribution, table of critical values,
826
t-distributions, 582–584, 826
t-procedures
pooled vs. unpoled, 629–630
as robust, 602, 607
sample size and, 568, 607–608, 629
skewness and outliers and, 602–609
slope and, 756
transformations and, 605–607
two samples vs. paired design,
640–647
t-tables, 565–566, 828
t-tests, 589–593
conditions for, 589, 820–821
paired data vs. single sample in,
640–641
use of, 640
t*-values
calculator technique, 566, 584
confidence intervals based on,
565–573, 618–623, 762–763
confidence level and, 566
defined, 565
degrees of freedom and, 566, 584, 621
outliers and, 604
P-values and, 593–594
significance testing and, 624–628,
760–761
t-distributions, 582–584, 828
t-tables, 565–566, 828
z-values compared to, 583
for proportions, 493–494, 499–500
for slope, 755–756, 758, 760–761, 790
for surveys, 531, 532
symbol for, 494
as t-statistic, 631
use of, 499
as z-score, 494
See also t-procedures; z-score

testing, samples and, 219
tests of significance. See significance tests
theoretical probability, probability distributions from, 361–364
third quartile. See upper quartile
time as dependent data, 765
and patterns of data, 179, 193
plots of, 822
transformations and, 184, 193
transformations, 179, 193
calculator technique, 192
exponential functions and. See exponential functions
heteroscedasticity and, 774–775, 778
linear vs. nonlinear, 179
log-log, 187–189, 774–779
logarithmic. See logarithmic transformations
of paired data, 648–649
power. See power transformations
recentering. See recentering
reciprocal transformations, 190, 605–607
regression lines and, 189, 774–779, 781
rescaling. See rescaling
of skewness and outliers, data with, 605–607, 648–649
square root transformations, 649–650
treatments
balanced design and, 266
defined, 244
variability between vs. within, 262–263, 270–272
tree diagrams, 294–295, 327–328
trend, 107, 108
trimmed mean, 73
truthfulness, surveys and, 226–227
two-stage sample, 238, 240
two-way tables, 692–693, 715
Type I error, 501–503, 506, 593
Type II error, 501–506, 594

U
uniform (rectangular) distributions, 28–30, 37
unimodal distributions, 35
units, experimental, 252–255
blocks, 267, 268–270, 273, 274
defined, 252
units of measurement
calculator technique exploring changes in, 76
correlation as quantity without, 143
estimation and, 50
standardizing. See z-score summary statistics and changes in, 75–76
units of population, 219
univariate data
categorical, 805
measurement, 807
unknown percentage problems, 84
recentering and rescaling and, 83
solving, 87–90
unknown values problems, 84–85
recentering and rescaling and, 83
solving, 87–90
upper quartile
defined, 34
in five-number summary, 61
interquartile range and, 59

V
variability
as challenge in data analysis, 4
interquartile range and, 59
linear models and, 741
management of, 262–263, 269, 270–274
point estimators and, 415, 417
tslope and, 739–749
variables
bivariate. See bivariate quantitative data
categorical. See categorical variables
defined, 4
identification of, 107
lurking, 107, 109, 147, 243–244
predictor (explanatory), 120
quantitative. See quantitative variables
random, 361, 372–373
response (predicted), 120
variance
calculator technique, 367
defined, 65
independent selection and, 372–373, 445
pooling and, 630
of a probability distribution, 365–367, 371–375
sampling distributions of the sum and difference and, 445
variation
coefficient of determination and, 148–151
random, 739
in sampling, 491
varying strength, 107
vertical axis, of histograms, 45
voluntary response bias, 224, 246

W
waiting-time problems, 393
weak association, 107, 108
See also correlation; strength of association
Westvaco case study, 3–8, 11–16, 817–822
whiskers, 61, 62

X
x-axis, 83

Y
y-axis, histograms and, 45
y-intercept, symbols for (b₀, β₀), 125, 739
See also slope-intercept form

Z
z-score, 83
calculator technique, 84, 85, 89
chi-square compared to, 685, 702–705
comparisons using, 87
computation of, 86
correlation and, 142–143
defined, 86
finding value with known, 86
confidence intervals using, 476–477
test statistic as, 494
unstandardizing, 86
using z-table, 84–85
z*. See critical values
Photo Credits

Chapter 1
2: John Olson/Corbis. 4: Zigy Kaluzny/Getty Images. 10: Jeff Greenberg/The Image Works.

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

Chapter 8

Chapter 9

Chapter 10

Chapter 11

Chapter 12