EXAMPLE 1.9 Pulling wood apart

Student engineers learn that, although handbooks give the strength of a material as a single number, in fact the strength varies from piece to piece. A vital lesson in all fields of study is that "variation is everywhere." Here are data from a typical student laboratory exercise: the load in pounds needed to pull apart pieces of Douglas fir 4 inches long and 1.5 inches square.

33,190  31,860  32,590  26,520  33,280
32,320  33,020  32,030  30,460  32,700
23,040  30,930  32,720  33,650  32,340
24,050  30,170  31,300  28,730  31,920

A stemplot of these data would have very many stems and no leaves or just one leaf on most stems. So we first round the data to the nearest hundred pounds. The rounded data are

332  319  326  265  333  323  330  320  305  327
230  309  327  337  323  241  302  313  287  319

FIGURE 1.10
Stemplot of the percents of foreign-born residents in the states, for Example 1.8. Each stem is a percent and leaves are tenths of a percent.
FIGURE 1.11
Stemplot of the breaking strength of pieces of wood, rounded to the nearest hundred pounds, for Example 1.9. Stems are thousands of pounds and leaves are hundreds of pounds.

Now we can make a stemplot with the first two digits (thousands of pounds) as stems and the third digit (hundreds of pounds) as leaves. Figure 1.11 is the stemplot. Rotate the stemplot counterclockwise so that it resembles a histogram, with 230 at the left end of the scale. This makes it clear that the distribution is skewed to the left. The midpoint is around 320 (32,000 pounds) and the spread is from 230 to 337. Because of the strong skew, we are reluctant to call the smallest observations outliers. They appear to be part of the long left tail of the distribution. Before using wood like this in construction, we should ask why some pieces are much weaker than the rest.

Comparing Figures 1.10 (right-skewed) and 1.11 (left-skewed) reminds us that the direction of skewness is the direction of the long tail, not the direction where most observations are clustered.

You can also split stems in a stemplot to double the number of stems when all the leaves would otherwise fall on just a few stems. Each stem then appears twice. Leaves 0 to 4 go on the upper stem, and leaves 5 to 9 go on the lower stem. If you split the stems in the stemplot of Figure 1.11, for example, the 32 and 33 stems become

Rounding and splitting stems are matters for judgment, like choosing the classes in a histogram. The wood strength data require rounding but don’t require splitting stems. The One-Variable Statistical Calculator applet on the text CD and Web site allows you to decide whether to split stems, so that it is easy to see the effect.

APPLY YOUR KNOWLEDGE

1.10 Older Americans. Make a stemplot of percent of residents aged 65 years and over in each of the 50 states and the District of Columbia in Table 1.2. Use whole percents as your stems. Because the stemplot preserves the actual value of the
1.20 The shape of the distribution in Exercise 1.19 is
(a) clearly skewed to the right.
(b) roughly symmetric.
(c) clearly skewed to the left.

1.21 The center of the distribution in Exercise 1.19 is
close to
(a) 22 minutes.  (b) 23.4 minutes.
(c) 15.5 to 30.9 minutes.

1.22 You look at real estate ads for houses in Naples,
Florida. There are many houses ranging from $200,000 to
$500,000 in price. The few houses on the water, however,
have prices up to $15 million. The distribution of house
prices will be
(a) skewed to the left.
(b) roughly symmetric.
(c) skewed to the right.

1.23 Medical students. Students who have finished medi-
cal school are assigned to residencies in hospitals to receive
further training in a medical specialty. Here is part of a hypo-
thesis data base of students seeking residency positions.
USMLE is the student’s score on Step 1 of the national medi-
cal licensing examination.

<table>
<thead>
<tr>
<th>Name</th>
<th>Medical school</th>
<th>Sex</th>
<th>Age</th>
<th>USMLE</th>
<th>Specialty sought</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abrams, Laurie</td>
<td>Florida</td>
<td>F</td>
<td>28</td>
<td>238</td>
<td>Family medicine</td>
</tr>
<tr>
<td>Brown, Gordon</td>
<td>Meharry</td>
<td>M</td>
<td>25</td>
<td>205</td>
<td>Radiology</td>
</tr>
<tr>
<td>Cabrera, Maria</td>
<td>Tufts</td>
<td>F</td>
<td>26</td>
<td>191</td>
<td>Pediatrics</td>
</tr>
<tr>
<td>Ismael, Miranda</td>
<td>Indiana</td>
<td>F</td>
<td>32</td>
<td>245</td>
<td>Internal medicine</td>
</tr>
</tbody>
</table>

(a) What individuals does this data set describe?
(b) In addition to the student’s name, how many variables does the
data set contain? Which of these variables are categorical and which
are quantitative?

1.24 Protecting wood. How can we help wood surfaces resist weath-
ering, especially when restoring historic wooden buildings? In a study of
this question, researchers prepared

1.25 What color is your car? The most popular colors for cars and light trucks vary with region and over time. In
North America white remains the top color choice, with
black the top choice in Europe and silver the top choice in
South America. Here is the distribution of the top colors for
vehicles sold globally in 2010:

<table>
<thead>
<tr>
<th>Color</th>
<th>Popularity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silver</td>
<td>25%</td>
</tr>
<tr>
<td>Black</td>
<td>24%</td>
</tr>
<tr>
<td>White</td>
<td>16%</td>
</tr>
<tr>
<td>Gray</td>
<td>16%</td>
</tr>
<tr>
<td>Red</td>
<td>6%</td>
</tr>
<tr>
<td>Blue</td>
<td>5%</td>
</tr>
<tr>
<td>Beige, brown</td>
<td>3%</td>
</tr>
<tr>
<td>Other colors</td>
<td></td>
</tr>
</tbody>
</table>

Fill in the percent of vehicles that are in other colors. Make a
graph to display the distribution of color popularity.
1.26 Facebook and MySpace audience. Although most social-networking Web sites in the United States have fairly short histories, the growth of these sites has been exponential. By far, the two most visited social-networking sites are Facebook.com and MySpace.com. Here is the age distribution of the audience for the two sites in December 2009:16

<table>
<thead>
<tr>
<th>Age group</th>
<th>Facebook visitors</th>
<th>MySpace visitors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under 25 years</td>
<td>26.8%</td>
<td>44.4%</td>
</tr>
<tr>
<td>25 to 34 years</td>
<td>23.0%</td>
<td>22.7%</td>
</tr>
<tr>
<td>35 to 49 years</td>
<td>31.6%</td>
<td>23.5%</td>
</tr>
<tr>
<td>Over 49 years</td>
<td>18.7%</td>
<td>9.4%</td>
</tr>
</tbody>
</table>

(a) Draw a bar graph for the age distribution of Facebook visitors. The leftmost bar should correspond to “under 25,” the next bar to “25 to 34,” and so on. Do the same for MySpace, using the same scale for the percent axis.
(b) Describe the most important difference in the age distribution of the audience for Facebook and MySpace. How does this difference show up in the bar graphs? Do you think it was important to order the bars by age to make the comparison easier?
(c) Explain why it is appropriate to use a pie chart to display either of these distributions. Draw a pie chart for each distribution. Do you think it is easier to compare the two distributions with bar graphs or pie charts? Explain your reasoning.

1.27 Deaths among young people. Among persons aged 15 to 24 years in the United States, the leading causes of death and number of deaths in 2008 were: accidents, 14,020; homicide, 5285; suicide, 4297; cancer, 1659; heart disease, 1059; congenital defects, 466.17

(a) Make a bar graph to display these data.
(b) To make a pie chart, you need one additional piece of information. What is it?

1.28 Hispanic origins. Figure 1.13 is a pie chart prepared by the U.S. Census Bureau to show the origin of the more than 43 million Hispanics in the United States in 2006.18

About what percent of Hispanics are Mexican? Puerto Rican? You see that it is hard to determine numbers from a pie chart. Bar graphs are much easier to use. (The U.S. Census Bureau did include the percents in its pie chart.)

1.29 Canadian students rate their universities. The National Survey of Student Engagement asked students at many universities, “How would you evaluate your entire educational experience at this university?” Here are the percents of senior-year students at Canada’s 10 largest primarily English-speaking universities who responded “Excellent”:19

<table>
<thead>
<tr>
<th>University</th>
<th>Excellent rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toronto</td>
<td>21%</td>
</tr>
<tr>
<td>York</td>
<td>18%</td>
</tr>
<tr>
<td>Alberta</td>
<td>23%</td>
</tr>
<tr>
<td>Ottawa</td>
<td>11%</td>
</tr>
<tr>
<td>Western Ontario</td>
<td>38%</td>
</tr>
<tr>
<td>British Columbia</td>
<td>18%</td>
</tr>
<tr>
<td>Calgary</td>
<td>14%</td>
</tr>
<tr>
<td>McGill</td>
<td>26%</td>
</tr>
<tr>
<td>Waterloo</td>
<td>36%</td>
</tr>
<tr>
<td>Concordia</td>
<td>21%</td>
</tr>
</tbody>
</table>

(a) The list is arranged in order of undergraduate enrollment. Make a bar graph with the bars in order of student rating.
(b) Explain carefully why it is not correct to make a pie chart of these data.

1.30 Do adolescent girls eat fruit? We all know that fruit is good for us. Many of us don’t eat enough. Figure 1.14 is a histogram of the number of servings of fruit per day claimed by 74 seventeen-year-old girls in a study in Pennsylvania.20

Describe the shape, center, and spread of this distribution.
CHAPTER 1  Picturing Distributions with Graphs

We read that IQ scores for large populations are centered at 100. What percent of these 78 students have scores above 100?

1.32 Returns on common stocks. The return on a stock is the change in its market price plus any dividend payments made. Total return is usually expressed as a percent of the beginning price. Figure 1.16 is a histogram of the distribution of the monthly returns for all stocks listed on U.S. markets from January 1985 to November 2010 (311 months). The extreme low outlier is the market crash of October 1987, when stocks lost 23% of their value in one month. The other two low outliers are 16% during August 1998, a month when the Dow Jones Industrial Average experienced its second largest drop in history to that time, and the financial crisis in October 2008 when stocks lost 17% of their value.

1.31 IQ test scores. Figure 1.15 is a stemplot of the IQ test scores of 78 seventh-grade students in a rural midwestern school.

(a) Four students had low scores that might be considered outliers. Ignoring these, describe the shape, center, and spread of the remainder of the distribution.

(b) We often read that IQ scores for large populations are centered at 100. What percent of these 78 students have scores above 100?

1.32 Returns on common stocks. The return on a stock is the change in its market price plus any dividend payments made. Total return is usually expressed as a percent of the beginning price. Figure 1.16 is a histogram of the distribution of the monthly returns for all stocks listed on U.S. markets from January 1985 to November 2010 (311 months). The extreme low outlier is the market crash of October 1987, when stocks lost 23% of their value in one month. The other two low outliers are 16% during August 1998, a month when the Dow Jones Industrial Average experienced its second largest drop in history to that time, and the financial crisis in October 2008 when stocks lost 17% of their value.

1.33 Name that variable. A survey of a large college class asked the following questions:

1. Are you female or male? (In the data, male = 0, female = 1.)
1.34 Food oils and health. Fatty acids, despite their unpleasant name, are necessary for human health. Two types of essential fatty acids, called omega-3 and omega-6, are not produced by our bodies and so must be obtained from our food. Food oils, widely used in food processing and cooking, are major sources of these compounds. There is some evidence that a healthy diet should have more omega-3 than omega-6. Table 1.4 (on the following page) gives the ratio of omega-3 to omega-6 in some common food oils. Values greater than 1 show that an oil has more omega-3 than omega-6.

(a) Make a histogram of these data, using classes bounded by the whole numbers from 0 to 6.
(b) What is the shape of the distribution? How many of the 30 food oils have more omega-3 than omega-6? What does this distribution suggest about the possible health effects of modern food oils?
(c) Table 1.4 contains entries for several fish oils (cod, herring, menhaden, salmon, sardine). How do these values support the idea that eating fish is healthy?

### Table 1.4 Omega-3 fatty acids as a fraction of omega-6 fatty acids in food oils

<table>
<thead>
<tr>
<th>OIL</th>
<th>RATIO</th>
<th>OIL</th>
<th>RATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perilla</td>
<td>5.33</td>
<td>Flaxseed</td>
<td>3.56</td>
</tr>
<tr>
<td>Walnut</td>
<td>0.20</td>
<td>Canola</td>
<td>0.46</td>
</tr>
<tr>
<td>Wheat germ</td>
<td>0.13</td>
<td>Soybean</td>
<td>0.13</td>
</tr>
<tr>
<td>Mustard</td>
<td>0.38</td>
<td>Grape seed</td>
<td>0.00</td>
</tr>
<tr>
<td>Sardine</td>
<td>2.16</td>
<td>Menhaden</td>
<td>1.95</td>
</tr>
<tr>
<td>Salmon</td>
<td>2.50</td>
<td>Herring</td>
<td>2.67</td>
</tr>
<tr>
<td>Mayonnaise</td>
<td>0.06</td>
<td>Soybean, hydrogenated</td>
<td>0.07</td>
</tr>
<tr>
<td>Cod liver</td>
<td>2.0C</td>
<td>Rice bran</td>
<td>0.05</td>
</tr>
<tr>
<td>Shortening</td>
<td>0.11</td>
<td>Butter</td>
<td>0.64</td>
</tr>
<tr>
<td>(household)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shortening</td>
<td>0.06</td>
<td>Sunflower</td>
<td>0.03</td>
</tr>
<tr>
<td>(industrial)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Margarine</td>
<td>0.05</td>
<td>Corn</td>
<td>0.01</td>
</tr>
<tr>
<td>Olive</td>
<td>0.06</td>
<td>Sesame</td>
<td>0.01</td>
</tr>
<tr>
<td>Shea nut</td>
<td>0.06</td>
<td>Cottonseed</td>
<td>0.00</td>
</tr>
<tr>
<td>Sunflower (oleic)</td>
<td>0.05</td>
<td>Palm</td>
<td>0.02</td>
</tr>
<tr>
<td>Sunflower (linoleic)</td>
<td>0.00</td>
<td>Cocoa butter</td>
<td>0.04</td>
</tr>
</tbody>
</table>

1.35 Where are the nurses? Table 1.5 (on the following page) gives the number of active nurses per 100,000 people in each state.24 Nurses

(a) Why is the number of nurses per 100,000 people a better measure of the availability of nurses than a simple count of the number of nurses in a state?

Figure 1.17 shows histograms of the student responses, in scrambled order and without scale markings. Which histogram goes with each variable? Explain your reasoning.

2. Are you right-handed or left-handed? (In the data, right = 0, left = 1.)
3. What is your height in inches?
4. How many minutes do you study on a typical weeknight?
TABLE 1.6 Annual carbon dioxide emissions in 2007 (metric tons per person)

<table>
<thead>
<tr>
<th>COUNTRY</th>
<th>CO2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afghanistan</td>
<td>0.0272</td>
</tr>
<tr>
<td>Algeria</td>
<td>4.1384</td>
</tr>
<tr>
<td>Argentina</td>
<td>4.6525</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>0.2773</td>
</tr>
<tr>
<td>Brazil</td>
<td>1.9373</td>
</tr>
<tr>
<td>Canada</td>
<td>16.9171</td>
</tr>
<tr>
<td>China</td>
<td>4.9194</td>
</tr>
<tr>
<td>Colombia</td>
<td>1.4301</td>
</tr>
<tr>
<td>Congo</td>
<td>0.0389</td>
</tr>
<tr>
<td>Egypt</td>
<td>2.3065</td>
</tr>
<tr>
<td>Ethiopia</td>
<td>0.0828</td>
</tr>
<tr>
<td>France</td>
<td>6.0207</td>
</tr>
<tr>
<td>Germany</td>
<td>9.5690</td>
</tr>
</tbody>
</table>

Make a stemplot to display the distribution of pups born per year. (Round to the nearest whole number and split the stems.) Describe the shape, center, and spread of the distribution. Are there any outliers?

1.38 Do women study more than men? We asked the students in a large first-year college class how many minutes they studied on a typical weeknight. Here are the responses of random samples of 30 women and 30 men from the class:

(a) Examine the data. Why are you not surprised that most responses are multiples of 10 minutes? What is the other common multiple found in the data? We eliminated one student who claimed to study 10,000 minutes per night. Are there any other responses you consider suspicious?

(b) Make a back-to-back stemplot to compare the two samples. That is, use one set of stems with two sets of leaves, one to the right and one to the left of the stems. (Draw a line on either side of the stems to separate stems and leaves.) Order both sets of leaves from smallest at the stem to largest away from the stem. Report the approximate midpoints of both groups. Does it appear that women study more than men (or at least claim that they do)?

**TABLE 1.6**

<table>
<thead>
<tr>
<th>WOMEN</th>
<th>MEN</th>
</tr>
</thead>
<tbody>
<tr>
<td>270</td>
<td>120</td>
</tr>
<tr>
<td>150</td>
<td>120</td>
</tr>
<tr>
<td>120</td>
<td>180</td>
</tr>
<tr>
<td>180</td>
<td>360</td>
</tr>
<tr>
<td>150</td>
<td>180</td>
</tr>
<tr>
<td>150</td>
<td>120</td>
</tr>
<tr>
<td>240</td>
<td>60</td>
</tr>
<tr>
<td>200</td>
<td>60</td>
</tr>
<tr>
<td>90</td>
<td>45</td>
</tr>
<tr>
<td>45</td>
<td>200</td>
</tr>
<tr>
<td>150</td>
<td>30</td>
</tr>
<tr>
<td>90</td>
<td>120</td>
</tr>
<tr>
<td>75</td>
<td>30</td>
</tr>
<tr>
<td>90</td>
<td>60</td>
</tr>
<tr>
<td>60</td>
<td>30</td>
</tr>
<tr>
<td>240</td>
<td>150</td>
</tr>
<tr>
<td>150</td>
<td>60</td>
</tr>
<tr>
<td>0</td>
<td>200</td>
</tr>
<tr>
<td>120</td>
<td>120</td>
</tr>
<tr>
<td>180</td>
<td>180</td>
</tr>
</tbody>
</table>

1.39 Fur seals on St. George Island. Make a time plot of the number of fur seals born per year from Exercise 1.37. What does the time plot show that your stemplot in Exercise 1.37 did not show? When you have data collected over time, a time plot is often needed to understand what is happening.
2010 (note that no data were acquired in 1995). To get a better feel for the magnitude of the numbers, the area of North America is approximately 24.5 million square kilometers (km²).

The two parts of this exercise will have you draw two graphs of these data.

(a) First make a time plot of the data. The severity of the ozone hole will vary from year to year depending on the meteorology of the atmosphere above Antarctica. Does the time plot illustrate only year-to-year variation or are there other patterns apparent? Specifically, is there a trend over any period of years? What about cyclical fluctuation? Explain in words the change in the average size of the ozone hole over this 30-year period.

(b) Now make a stemplot of the data. What is the midpoint of the distribution of ozone hole size? Do you think that the stemplot and the midpoint are a good description of this data set? Is there important information in the time plot that is not contained in the stemplot? When data are collected over time, you should always make a time plot.

1.46 To split or not to split. The data sets in the One-Variable Statistical Calculator applet on the text CD and Web site include the “pulling wood apart” data from Example 1.9. The applet rounds the data in the same way as Figure 1.11 (page 22). Use the applet to make a stemplot with split stems. Do you prefer this stemplot or that in Figure 1.11? Explain your choice.

**EXPLORING THE WEB**

1.47 Natural Gas Prices. The Department of Energy Web site contains information about monthly wholesale and retail prices for natural gas in each state. Go to www.eia.doe.gov/naturalgas/data.cfm and then click on the link Monthly Wholesale and Retail Prices. Under Area, choose a state of interest to you, make sure the Period is monthly, and then under Residential Price click on View History. A window will open with a time plot covering approximately a 20-year period, along with a table of the monthly residential prices for each year.

(a) If you have access to statistical software, you should use the Download Data (XLS file) link to save the data as an Excel (.xls) file on your computer. Then enter the data into your software package, and reproduce the time series plot using the graphical capabilities of your software package. Be sure you use an appropriate title and axis labels. If you do not have access to appropriate software, provide a rough sketch of the time plot that is given on the Web site.

(b) Is there a regular pattern of seasonal variation that repeats each year? Describe it. Are the prices increasing over time?

1.48 Hank Aaron’s home run record. The all-time home run leader prior to 2007 was Hank Aaron. You can find his career statistics by going to the Web site www.baseball-reference.com and then clicking on the Players tab at the top of the page and going to Hank Aaron.
information in your data can be described by a few numbers. These numerical summaries can be useful for describing a single distribution as well as for comparing the distributions from several groups of observations.

Two important features of a distribution are the center and the spread. For distributions that are approximately symmetric without outliers, the mean and standard deviation are important numeric summaries for describing and comparing distributions. But if the distribution is not symmetric and/or has outliers, the five-number summary often provides a better description.

The boxplot gives a picture of the five-number summary that is useful for a simple comparison of several distributions. Remember that the boxplot is based only on the five-number summary and does not have any information beyond these five numbers. Certain features of a distribution that are revealed in histograms and stemplots will not be evident from a boxplot alone. These include gaps in the data and the presence of several peaks. You must be careful when reducing a distribution to a few numbers to make sure that important information has not been lost in the process.

### CHECK YOUR SKILLS

2.15 The respiratory system can be a limiting factor in maximal exercise performance. Researchers from the United Kingdom studied the effect of two breathing frequencies on both performance times and several physiological parameters in swimming. Subjects were 10 male collegiate swimmers. Here are their times in seconds to swim 200 meters at 90% of race pace when breathing every second stroke in front-crawl swimming:

- 151.6
- 165.1
- 159.2
- 163.5
- 174.8
- 173.2
- 177.6
- 174.3
- 164.1
- 171.4

The mean of these data is
(a) 165.10.  (b) 167.48.  (c) 168.25.

2.16 The median of the data in Exercise 2.15 is
(a) 167.48.  (b) 168.25.  (c) 174.00.

2.17 The five-number summary of the data in Exercise 2.15 is
(a) 151.6, 159.2, 167.48, 174.8, 177.6.
(b) 151.6, 163.5, 168.15, 174.3, 177.6.
(c) 151.6, 159.2, 168.15, 174.8, 177.6.

2.18 If a distribution is skewed to the right,
(a) the mean is less than the median.
(b) the mean and median are equal.
(c) the mean is greater than the median.

2.19 What percent of the observations in a distribution lie between the first quartile and the third quartile?
(a) 25%  (b) 50%  (c) 75%

2.20 To make a boxplot of a distribution, you must know
(a) all of the individual observations.
(b) the mean and the standard deviation.
(c) the five-number summary.

2.21 The standard deviation of the 10 swim times in Exercise 2.15 (use your calculator) is about
(a) 7.4.  (b) 7.8.  (c) 8.2.

2.22 What are all the values that a standard deviation $s$ can possibly take?
(a) $0 \leq s$  (b) $0 \leq s \leq 1$  (c) $-1 \leq s \leq 1$

2.23 The correct units for the standard deviation in Exercise 2.21 are
(a) no units—it’s just a number.
(b) seconds.
(c) seconds squared.

2.24 Which of the following is least affected if an extreme high outlier is added to your data?
(a) The median
(b) The mean
(c) The standard deviation
2.25 Incomes of college grads. According to the Census Bureau's 2010 Current Population Survey, the mean and median 2009 income of people at least 25 years old who had a bachelor's degree but no higher degree were $46,931 and $38,762. Which of these numbers is the mean and which is the median? Explain your reasoning.

2.26 Saving for retirement. Retirement seems a long way off and we need money now, so saving for retirement is hard. Once every three years, the Board of Governors of the Federal Reserve System collects data on household assets and liabilities through the Survey of Consumer Finances (SCF). The most recent such survey was conducted in 2007, and the survey results were released to the public in April 2009. The survey presents data on household ownership of, and balances in, retirement savings accounts. Only 53.6% of households own retirement accounts. The mean value per household is $148,579, but the median value is just $45,000. For households in which the head of household is under 35, 42.6% own retirement accounts, the mean is $25,279, and the median is $9600. What explains the differences between the two measures of center, both for all households and for the under-35 age group?

2.27 University endowments. The National Association of College and University Business Officers collects data on college endowments. In 2009, 842 colleges and universities reported the value of their endowments. When the endowment values are arranged in order, what are the locations of the median and the quartiles in this ordered list?

2.28 Pulling wood apart. Example 1.9 (page 21) gives the breaking strengths of 20 pieces of Douglas fir.

(a) Give the five-number summary of the distribution of breaking strengths. (The stemplot, Figure 1.11, helps because it arranges the data in order, but you should use the unrounded values in numerical work.)

(b) The stemplot shows that the distribution is skewed to the left. Does the five-number summary show the skew? Remember that only a graph gives a clear picture of the shape of a distribution.

2.29 Comparing tropical flowers. An alternative presentation of the flower length data in Table 2.1 reports the five-number summary and uses boxplots to display the distributions. Do this. Do the boxplots fail to reveal any important information visible in the stemplots in Figure 2.4?

2.30 How much fruit do adolescent girls eat? Figure 1.14 (page 30) is a histogram of the number of servings of fruit per day claimed by 74 seventeen-year-old girls.

(a) With a little care, you can find the median and the quartiles from the histogram. What are these numbers? How did you find them?

(b) With a little care, you can also find the mean number of servings of fruit claimed per day. First use the information in the histogram to compute the sum of the 74 observations, and then use this to compute the mean. What is the relationship between the mean and median? Is this what you expected?

2.31 Guinea pig survival times. Here are the survival times in days of 72 guinea pigs after they were injected with infectious bacteria in a medical experiment. Survival times, whether of machines under stress or cancer patients after treatment, usually have distributions that are skewed to the right.

<table>
<thead>
<tr>
<th>Survival Times (days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>43 45 53 56 57 58 66 67 73</td>
</tr>
<tr>
<td>74 79 80 81 81 82 83</td>
</tr>
<tr>
<td>84 88 89 91 91 92 92 97 99 99</td>
</tr>
<tr>
<td>100 101 102 102 103 104 107 108</td>
</tr>
<tr>
<td>109 113 114 118 121 123 126 128 137 138</td>
</tr>
<tr>
<td>139 144 145 147 156 162 174 178 179 184</td>
</tr>
<tr>
<td>191 198 211 214 243 249 329 380 403 511</td>
</tr>
<tr>
<td>522 598</td>
</tr>
</tbody>
</table>

(a) Graph the distribution and describe its main features. Does it show the expected right-skew?

(b) Which numerical summary would you choose for these data? Calculate your chosen summary. How does it reflect the skewness of the distribution?

(a) In the context of this study, what do the negative values in the data set mean?
(b) Give a graphical comparison of the weight loss distribution for both groups using side-by-side boxplots. Provide appropriate numerical summaries for the two distributions and identify any high outliers in either group. What can you say about the effects of gastric banding versus lifestyle intervention on weight loss for the subjects in this study?
(c) The measured variable was weight loss in kilograms. Would two subjects with the same weight loss always have similar benefits from a weight reduction program? Does it depend on their initial weights? Other variables considered in this study were the percent of excess weight lost and the reduction in BMI. Do you see any advantages to either of these variables when comparing weight loss for two groups?
(d) One subject from the gastric-banding group dropped out of the study and seven subjects from the lifestyle group dropped out. Of the seven dropouts in the lifestyle group, six had gained weight at the time they dropped out. If all subjects had completed the study, how do you think it would have affected the comparison between the two groups?

Exercises 2.44 to 2.49 ask you to analyze data without having the details outlined for you. The exercise statements give you the State step of the four-step process. In your work, follow the Plan, Solve, and Conclude steps as illustrated in Example 2.9.

2.44 Athletes' salaries. The Montreal Canadiens were founded in 1909 and are the longest continuously operating professional ice hockey team. They have won 24 Stanley Cups, making them one of the most successful professional sports teams of the traditional four major sports of Canada and the United States. Table 2.2 gives the salaries of the 2010—2011 roster. Provide the team owner with a full description of the distribution of salaries and a brief summary of its most important features.

![Montreal Canadiens](image)

<table>
<thead>
<tr>
<th>TABLE 2.2</th>
<th>Salaries for the 2010—2011 Montreal Canadiens</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLAYER</td>
<td>SALARY</td>
</tr>
<tr>
<td>Scott Gomez</td>
<td>$6,000,000</td>
</tr>
<tr>
<td>Mike Cammalleri</td>
<td>$5,000,000</td>
</tr>
<tr>
<td>Jaroslav Spacek</td>
<td>$3,833,000</td>
</tr>
<tr>
<td>Carey Price</td>
<td>$2,500,000</td>
</tr>
<tr>
<td>Benoit Pouliot</td>
<td>$1,350,000</td>
</tr>
<tr>
<td>Max Pacioretty</td>
<td>$875,000</td>
</tr>
<tr>
<td>Yannick Weber</td>
<td>$637,500</td>
</tr>
<tr>
<td>David Desharnais</td>
<td>$550,000</td>
</tr>
</tbody>
</table>

2.45 Returns on stocks. How well have stocks done over the past generation? The Wilshire 5000 index describes the average performance of all U.S. stocks. The average is weighted by the total market value of each company's stock, so think of the index as measuring the performance of the average investor. Page 64 gives the percent returns on the Wilshire 5000 index for the years from 1971 to 2010:
### TABLE 2.5 Earnings (dollars) of a sample of 200 Canadians in 1901

<table>
<thead>
<tr>
<th>Amount</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>400</td>
<td>360</td>
</tr>
<tr>
<td>250</td>
<td>525</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
</tr>
<tr>
<td>750</td>
<td>300</td>
</tr>
<tr>
<td>1000</td>
<td>150</td>
</tr>
<tr>
<td>200</td>
<td>600</td>
</tr>
<tr>
<td>300</td>
<td>150</td>
</tr>
<tr>
<td>400</td>
<td>120</td>
</tr>
<tr>
<td>500</td>
<td>150</td>
</tr>
<tr>
<td>600</td>
<td>150</td>
</tr>
<tr>
<td>700</td>
<td>150</td>
</tr>
<tr>
<td>800</td>
<td>150</td>
</tr>
<tr>
<td>900</td>
<td>150</td>
</tr>
<tr>
<td>1000</td>
<td>150</td>
</tr>
</tbody>
</table>

The United States and several other countries appear to be high outliers.

(a) Give the five-number summary. Explain why this summary suggests that the distribution is right-skewed.

(b) Which countries are outliers according to the $1.5 \times IQR$ rule? Make a stemplot of the data or look at your stemplot from Exercise 1.36. Do you agree with the rule’s suggestions about which countries are and are not outliers?

#### 2.52 Athletes’ salaries.

Which members of the Montreal Canadiens (Table 2.1) have salaries that are suspected outliers by the $1.5 \times IQR$ rule?

#### 2.53 Canadians’ earnings in 1901.

The Canadians’ earnings in Exercise 2.49 are right-skewed, with a few large incomes. Which incomes are suspected outliers by the $1.5 \times IQR$ rule?

---

### EXPLORING THE WEB

#### 2.54 Home run leaders.

The three top players on the career home run list are Barry Bonds, Hank Aaron, and Babe Ruth. You can find their home run statistics by going to the Web site [www.baseball-reference.com](http://www.baseball-reference.com) and then clicking on the Players tab at the top of the page. Construct three side-by-side boxplots comparing the yearly home run production of Barry Bonds, Hank Aaron, and Babe Ruth. Describe any differences that you observe. It is worth noting that in his first four seasons, Babe Ruth was primarily a pitcher. If these four seasons are ignored, how does Babe Ruth compare with Barry Bonds and Hank Aaron?
these two numerical summaries give a complete description of the distribution of our data. It is important to remember that not all distributions can be well approximated by a Normal curve. In these cases, calculations based on the Normal distribution can be misleading.

Normal distributions are also good approximations to many kinds of chance outcomes such as the proportion of heads in many tosses of a coin (this setting will be described in more detail in Chapter 20). And when we discuss statistical inference in Part III of the text, we will find that many procedures based on Normal distributions work well for other roughly symmetric distributions.

3.15 Which of these variables is most likely to have a Normal distribution?
(a) Income per person for 150 different countries.
(b) Sale prices of 200 homes in a suburb of Chicago.
(c) Heights of 100 white pine trees in a forest.

3.16 To completely specify the shape of a Normal distribution, you must give
(a) the mean and the standard deviation.
(b) the five-number summary.
(c) the median and the quartiles.

3.17 Figure 3.14 shows a Normal curve. The mean of this distribution is
(a) 0. (b) 2. (c) 3.

3.18 The standard deviation of the Normal distribution in Figure 3.14 is
(a) 2. (b) 3. (c) 5.

3.19 The length of human pregnancies from conception to birth varies according to a distribution that is approximately Normal with mean 266 days and standard deviation 16 days. About 95% of all pregnancies last between
(a) 250 and 282 days. (b) 234 and 298 days.
(c) 218 and 314 days.

3.20 The scores of adults on an IQ test are approximately Normal with mean 100 and standard deviation 15. The organization MENSA, which calls itself "the high-IQ society," requires an IQ score of 130 or higher for membership. What percent of adults would qualify for membership?
(a) 95% (b) 5% (c) 2.5%

3.21 The scores of adults on an IQ test are approximately Normal with mean 100 and standard deviation 15. Clara scores 127 on such a test. Her z-score is about
(a) 1.27. (b) 1.80. (c) 8.47.

3.22 The proportion of observations from a standard Normal distribution that take values greater than 1.83 is about
(a) 0.0641. (b) 0.0359. (c) 0.0336.

3.23 The proportion of observations from a standard Normal distribution that take values less than −0.75 is about
(a) 0.2266. (b) 0.7734. (c) 0.8023.

3.24 The scores of adults on an IQ test are approximately Normal with mean 100 and standard deviation 15. Clara scores 127 on such a test. She scores higher than what percent of all adults?
(a) About 10% (b) About 90% (c) About 96%
CHAPTER 3 EXERCISES

3.25 **Understanding density curves.** Remember that it is areas under a density curve, not the height of the curve, that give proportions in a distribution. To illustrate this, sketch a density curve that has a tall, thin peak at 0 on the horizontal axis but has most of its area close to 1 on the horizontal axis without a high peak at 1.

3.26 **Daily activity.** It appears that people who are mildly obese are less active than leaner people. One study looked at the average number of minutes per day that people spend standing or walking. Among mildly obese people, minutes of activity varied according to the N(373, 67) distribution. Minutes of activity for lean people had the N(526, 107) distribution. Within what limits do the active minutes for about 95% of the people in each group fall? Use the 68–95–99.7 rule.

3.27 **Low IQ test scores.** Scores on the Wechsler Adult Intelligence Scale (WAIS) are approximately Normal with mean 100 and standard deviation 15. People with WAIS scores below 70 are considered mentally retarded when, for example, applying for Social Security disability benefits. According to the 68–95–99.7 rule, about what percent of adults are retarded by this criterion?

3.28 **Standard Normal drill.** Use Table A to find the proportion of observations from a standard Normal distribution that fall in each of the following regions. In each case, sketch a standard Normal curve and shade the area representing the region.

(a) \( z \leq -1.25 \)
(b) \( z \geq -1.25 \)
(c) \( z > 2.17 \)
(d) \( -1.25 < z < 2.17 \)

3.29 **Standard Normal drill.**

(a) Find the number \( z \) such that the proportion of observations that are less than \( z \) in a standard Normal distribution is 0.6.
(b) Find the number \( z \) such that 15% of all observations from a standard Normal distribution are greater than \( z \).

3.30 **Fruit flies.** The thorax lengths in a population of male fruit flies follow a Normal distribution with mean 0.800 millimeters (mm) and standard deviation 0.078 mm.

(a) What proportion of flies have thorax lengths less than 0.7 mm?
(b) What proportion have thorax lengths greater than 1 mm?
(c) What proportion have thorax lengths between 0.7 and 1 mm?

3.31 **Acid rain?** Emissions of sulfur dioxide by industry set off chemical changes in the atmosphere that result in "acid rain." The acidity of liquids is measured by pH on a scale of 0 to 14. Distilled water has pH 7.0, and lower pH values indicate acidity. Normal rain is somewhat acidic, so acid rain is sometimes defined as rainfall with a pH below 5.0. The pH of rain at one location varies among rainy days according to a Normal distribution with mean 5.43 and standard deviation 0.54. What proportion of rainy days have rainfall with pH below 5.0?

3.32 **Runners.** In a study of exercise, a large group of male runners walk on a treadmill for 6 minutes. Their heart rates in beats per minute at the end vary from runner to runner according to the N(104, 12.5) distribution. The heart rates for male nonrunners after the same exercise have the N(130, 17) distribution.
(a) What percent of the runners have heart rates above 130?
(b) What percent of the nonrunners have heart rates above 130?

3.33 **A milling machine.** Automated manufacturing operations are quite precise but still vary, often with distributions that are close to Normal. The width in inches of slots cut by a milling machine follows approximately the N(0.8750, 0.0012) distribution. The specifications allow slot widths between 0.8720 and 0.8780 inch. What proportion of slots meet these specifications?

3.34 **Body mass index.** Your body mass index (BMI) is your weight in kilograms divided by the square of your height in meters. Many online BMI calculators allow you to enter weight in pounds and height in inches. High BMI is a common but controversial indicator of overweight or obesity. A study by the National Center for Health Statistics found that the BMI of American young women (ages 20 to 29) is approximately Normal with mean 26.5 and standard deviation 6.4.

(a) People with BMI less than 18.5 are often classified as "underweight." What percent of young women are underweight by this criterion?
(b) People with BMI over 30 are often classified as "obese." What percent of young women are obese by this criterion?
Miles per gallon. In its Fuel Economy Guide for model year 2010 vehicles, the Environmental Protection Agency gives data on 1101 vehicles. There are a number of high outliers, mainly hybrid gas-electric vehicles. If we ignore the vehicles identified as outliers, however, the combined city and highway gas mileage of the other 1082 vehicles is approximately Normal with mean 20.3 miles per gallon (mpg) and standard deviation 4.3 mpg. Exercises 3.35 to 3.38 concern this distribution.

3.35 In my Chevrolet. The 2010 Chevrolet Camaro with an eight-cylinder engine and automatic transmission has a combined gas mileage of 19 mpg. What percent of all vehicles have better gas mileage than the Camaro?

3.36 The bottom 10%. How low must a 2010 vehicle’s gas mileage be in order to fall in the bottom 10% of all vehicles?

3.37 The middle half. The quartiles of any distribution are the values with cumulative proportions 0.25 and 0.75. They span the middle half of the distribution. What are the quartiles of the distribution of gas mileage?

3.38 Quintiles. The quintiles of any distribution are the values with cumulative proportions 0.20, 0.40, 0.60, and 0.80. What are the quintiles of the distribution of gas mileage?

3.39 What’s your percentile? Reports on a student’s ACT, SAT, or MCAT usually give the percentile as well as the actual score. The percentile is just the cumulative proportion stated as a percent: the percent of all scores that were lower than this one. In 2010, the total MCAT scores were close to Normal with mean 25.0 and standard deviation 6.4. William scored 32. What was his percentile?

3.40 Perfect SAT scores. It is possible to score higher than 1600 on the combined Mathematics and Reading portions of the SAT, but scores of 1600 and above are reported as 1600. The distribution of SAT scores (combining Mathematics and Reading) was close to Normal with mean 1021 and standard deviation 211. What proportion of SAT scores for these two parts were reported as 1600?

3.41 Heights of women. The heights of women aged 20 to 29 follow approximately the N(64.3, 2.7) distribution. Men the same age have heights distributed as N(69.9, 3.1). What percent of young women are taller than the mean height of young men?

3.42 Weights aren’t Normal. The heights of people of the same sex and similar ages follow a Normal distribution reasonably closely. Weights, on the other hand, are not Normally distributed. The weights of women aged 20 to 29 have mean 155.9 pounds and median 144.0 pounds. The first and third quartiles are 124.1 pounds and 173.7 pounds. What can you say about the shape of the weight distribution? Why?

3.43 A surprising calculation. Changing the mean and standard deviation of a Normal distribution by a moderate amount can greatly change the percent of observations in the tails. Suppose that a college is looking for applicants with SAT Math scores of 750 and above.

(a) In 2010, the scores of men on the SAT Math test followed the N(534, 118) distribution. What percent of men scored 750 or better?

(b) Women’s SAT Math scores that year had the N(500, 112) distribution. What percent of women scored 750 or better? You see that the percent of men above 750 is almost three times the percent of women with such high scores. Why this is true is controversial. (On the other hand, women score higher than men on the new SAT Writing test, though by a smaller amount.)

3.44 Grading managers. Some companies “grade on a bell curve” to compare the performance of their managers and professional workers. This forces the use of some low performance ratings so that not all workers are listed as “above average.” Ford Motor Company’s “performance management process” for this year assigned 10% A grades, 80% B grades, and 10% C grades to the company’s managers. Suppose that Ford’s performance scores really are Normally distributed. This year, managers with scores less than 25 received C’s and those with scores above 475 received A’s. What are the mean and standard deviation of the scores?

3.45 Osteoporosis. Osteoporosis is a condition in which the bones become brittle due to loss of minerals. To diagnose osteoporosis, an elaborate apparatus measures bone mineral density (BMD). BMD is usually reported in standardized form. The standardization is based on a population of healthy young adults. The World Health Organization (WHO) criterion for osteoporosis is a BMD 2.5 standard deviations below the mean for young adults. BMD measurements in a population of people similar in age and sex roughly follow a Normal distribution.

(a) What percent of healthy young adults have osteoporosis by the WHO criterion?

(b) Women aged 70 to 79 are of course not young adults. The mean BMD in this age is about –2 on the standard scale
for young adults. Suppose that the standard deviation is the same as for young adults. What percent of this older population have osteoporosis?

In later chapters we will meet many statistical procedures that work well when the data are "close enough to Normal." Exercises 3.46 to 3.50 concern data that are mostly close enough to Normal for statistical work, while Exercise 3.51 concerns data for which the data are not close to Normal. These exercises ask you to do data analysis and Normal calculations to investigate how close to Normal real data are.

### 3.46 Normal is only approximate: IQ test scores.
Here are the IQ test scores of 31 seventh-grade girls in a Midwest school district: 114, 100, 104, 89, 102, 91, 114, 114, 103, 105, 108, 130, 120, 132, 111, 128, 118, 119, 86, 72, 111, 103, 74, 112, 107, 103, 96, 96, 112, 112, 93. Are the IQ test scores of 31 seventh-grade girls in a Midwest school district Normal? What proportion of the scores are within one standard deviation of the mean? Within two standard deviations of the mean? What would these proportions be in an exactly Normal distribution?

(a) We expect IQ scores to be approximately Normal. Make a stemplot to check that there are no major departures from Normality.

(b) Nonetheless, proportions calculated from a Normal distribution are not always very accurate for small numbers of observations. Find the mean \( \bar{x} \) and standard deviation \( s \) for these IQ scores. What proportion of the scores are within one standard deviation of the mean? Within two standard deviations of the mean? What would these proportions be in an exactly Normal distribution?

### 3.47 Normal is only approximate: ACT scores.
Scores on the ACT test for the 2010 high school graduating class had mean 21.0 and standard deviation 5.2. In all, 1,568,835 students in this class took the test. Of these, 145,000 had scores higher than 28 and another 50,860 had scores exactly 28. ACT scores are always whole numbers. The exactly Normal \( N(21.0, 5.2) \) distribution can include any value, not just whole numbers. What is more, there is no area exactly above 28 under the smooth Normal curve. So ACT scores can be only approximately Normal. To illustrate this fact, find

(a) the percent of 2010 ACT scores greater than 28.

(b) the percent of 2010 ACT scores greater than or equal to 28.

(c) the percent of observations from the \( N(21.0, 5.2) \) distribution that are greater than 28. (The percent greater than or equal to 28 is the same, because there is no area exactly above 28.)

### 3.48 Are the data Normal? Acidity of rainfall.
Exercise 3.31 concerns the acidity (measured by pH) of rainfall. A sample of 105 rainwater specimens had mean pH 5.43, standard deviation 0.54, and five-number summary 4.33, 5.05, 5.44, 5.79, 6.81. Is the pH of rainwater specimens Normally distributed?

(a) Compare the mean and median and also the distances of the two quartiles from the median. Does it appear that the distribution is quite symmetric? Why?

(b) If the distribution is really \( N(5.43, 0.54) \), what proportion of observations would be less than 5.05? Less than 5.79? Do these proportions suggest that the distribution is close to Normal? Why?

### 3.49 Are the data Normal? SAT Critical Reading scores.
Georgia Southern University (GSU) had 2417 students with regular admission in their freshman class of 2010. For each student, data are available on their SAT and ACT scores, if taken, high school GPA, and the college within the university to which they were admitted. Here are the first 20 SAT Critical Reading scores from that data set: 650, 490, 580, 450, 570, 540, 510, 530, 510, 560, 560, 590, 470, 690, 530, 570, 460, 590, 530, 490.

The complete data set is on the text Web site and CD, which contains both the original scores and the ordered scores.

(a) Make a histogram of the distribution (if your software allows it, superimpose a Normal curve over the histogram as in Figure 3.1). Although the resulting histogram depends a bit on your choice of classes, the distribution appears roughly symmetric with no outliers.

(b) Find the mean, median, standard deviation, and quartiles for these data. Comparing the mean and median and comparing the distances of the two quartiles from the median suggest that the distribution is quite symmetric. Why?

(c) In 2010, the mean score on the Critical Reading portion of the SAT for all college-bound seniors was 501. If the distribution were exactly Normal with the mean and standard deviation you found in part (b), what proportion of regularly admitted GSU freshmen scored above the mean for all college-bound seniors?

(d) Compute the exact proportion of regularly admitted GSU freshmen who scored above the mean for all college-bound seniors. It will be simplest to use the ordered scores in the data file to calculate this. How does this percent compare with the percent calculated in part (c)? Despite the discrepancy, this distribution is "close enough to Normal" for statistical work in later chapters.

### 3.50 Are the data Normal? Monsoon rains.
Here are the amounts of summer monsoon rainfall (millimeters) for India in the 100 years from 1901 to 2000: 14.2, 12.3, 15.4, 16.5, 17.6, 18.7, 19.8, 20.9, 21.0, 22.1, 23.2, 24.3, 25.4, 26.5, 27.6, 28.7, 29.8, 30.9, 31.0, 32.1, 33.2, 34.3, 35.4, 36.5, 37.6, 38.7, 39.8, 40.9, 41.0, 42.1, 43.2, 44.3, 45.4, 46.5, 47.6, 48.7, 49.8, 50.9, 51.0, 52.1, 53.2, 54.3, 55.4, 56.5, 57.6, 58.7, 59.8, 60.9, 61.0, 62.1, 63.2, 64.3, 65.4, 66.5, 67.6, 68.7, 69.8, 70.9, 71.0, 72.1, 73.2, 74.3, 75.4, 76.5, 77.6, 78.7, 79.8, 80.9, 81.0, 82.1, 83.2, 84.3, 85.4, 86.5, 87.6, 88.7, 89.8, 90.9, 91.0, 92.1, 93.2, 94.3, 95.4, 96.5, 97.6, 98.7, 99.8, 100.0.

The complete data set is on the text Web site and CD, which contains both the original scores and the ordered scores.

(a) Make a histogram of the distribution (if your software allows it, superimpose a Normal curve over the histogram as in Figure 3.1). Although the resulting histogram depends a bit on your choice of classes, the distribution appears roughly symmetric with no outliers.

(b) Find the mean, median, standard deviation, and quartiles for these data. Comparing the mean and median and comparing the distances of the two quartiles from the median suggest that the distribution is quite symmetric. Why?

(c) In 2010, the mean score on the Critical Reading portion of the SAT for all college-bound seniors was 501. If the distribution were exactly Normal with the mean and standard deviation you found in part (b), what proportion of regularly admitted GSU freshmen scored above the mean for all college-bound seniors?

(d) Compute the exact proportion of regularly admitted GSU freshmen who scored above the mean for all college-bound seniors. It will be simplest to use the ordered scores in the data file to calculate this. How does this percent compare with the percent calculated in part (c)? Despite the discrepancy, this distribution is "close enough to Normal" for statistical work in later chapters.
summarized as either a positive association (high values of the two variables tend to occur together) or a negative association (high values of one variable tend to occur with low values of the other variable). Forms to watch for are straight-line patterns, curved patterns, or clusters. The strength of a relationship is determined by how close the points in the scatterplot lie to a simple form such as a straight line or curve.

One of the simplest forms is a linear relationship, where the points suggest a straight-line pattern. Correlation is a number that summarizes the strength and direction of a linear relation. Positive values of the correlation correspond to a positive association. Negative values correspond to a negative association. The closer the absolute value of the correlation is to 1, the stronger the linear relationship (the more closely the points in the scatterplot come to lying on a straight line).

It is tempting to assume that the patterns we observe in our data hold for values of our variables that we have not observed—in other words, that additional data would continue to conform to these patterns. The process of identifying underlying patterns would seem to assume that this is the case. But is this assumption justified? We will return to this issue in Part IV of the book.

### CHECK YOUR SKILLS

**4.14** You have data for many years on the average price of a barrel of oil and the average retail price of a gallon of unleaded regular gasoline. When you make a scatterplot, the explanatory variable on the x axis
(a) is the price of oil. (b) is the price of gasoline. (c) can be either oil price or gasoline price.

**4.15** In a scatterplot of the average price of a barrel of oil and the average retail price of a gallon of gasoline, you expect to see
(a) a positive association. (b) very little association. (c) a negative association.

**4.16** Figure 4.7 is a scatterplot of school GPA against IQ test scores for 15 seventh-grade students. There is one low outlier in the plot. The IQ and GPA scores for this student are
(a) IQ = 0.5, GPA = 103. (b) IQ = 103, GPA = 0.5. (c) IQ = 103, GPA = 7.6.

**4.17** If we leave out the low outlier, the correlation for the remaining 14 points in Figure 4.7 is closest to
(a) 0.9. (b) −0.9. (c) 0.1.

**4.18** What are all the values that a correlation $r$ can possibly take?
(a) $r \geq 0$ (b) $0 \leq r \leq 1$ (c) $-1 \leq r \leq 1$

**4.19** If the correlation between two variables is close to 0, you can conclude that a scatterplot would show
(a) a strong straight-line pattern. (b) a cloud of points with no visible pattern. (c) no straight-line pattern, but there might be a strong pattern of another form.

**4.20** The points on a scatterplot lie very close to a straight line. The correlation between x and y is close to
(a) −1. (b) 1. (c) either −1 or 1, we can’t say which.

**4.21** If men always married women who were two years younger than themselves, the correlation between the ages of husband and wife would be
(a) 1. (b) −1. (c) Can’t tell without seeing the data.

**4.22** For a biology project, you measure the weight in grams and the tail length in millimeters of a group of mice. The correlation is $r = 0.7$. If you had measured tail length in centimeters instead of millimeters, what would be the correlation? (There are 10 millimeters in a centimeter.)
(a) $0.7/10 = 0.07$ (b) 0.7 (c) $(0.7)(10) = 7$

**4.23** Because elderly people may have difficulty standing to have their heights measured, a study looked at predicting overall height from height to the knee. Here are data (in centimeters) for six elderly men:

<table>
<thead>
<tr>
<th>Knee height (cm)</th>
<th>Height (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>57.7</td>
<td>192.1</td>
</tr>
<tr>
<td>47.4</td>
<td>153.3</td>
</tr>
<tr>
<td>43.5</td>
<td>146.4</td>
</tr>
<tr>
<td>44.8</td>
<td>162.7</td>
</tr>
<tr>
<td>55.2</td>
<td>169.1</td>
</tr>
<tr>
<td>54.6</td>
<td>177.8</td>
</tr>
</tbody>
</table>

Use your calculator or software: the correlation between knee height and overall height is about
(a) $r = 0.08$. (b) $r = 0.89$. (c) $r = 0.74$. 
CHAPTER 4 EXERCISES

4.24 Scores at the Masters. The Masters is one of the four major golf tournaments. Figure 4.8 is a scatterplot of the scores for the first two rounds of the 2010 Masters for all the golfers entered. Only the 60 golfers with the lowest two-round total advance to the final two rounds. The plot has a grid pattern because golf scores must be whole numbers.

(a) Read the graph: What was the lowest score in the first round of play? How many golfers had this low score? What were their scores in the second round?
(b) Read the graph: Sandy Lyle had the highest score in the second round. What was this score? What was Lyle's score in the first round?
(c) Is the correlation between first-round scores and second-round scores closest to \( r = 0.1 \), \( r = 0.5 \), or \( r = 0.9 \)? Explain your choice. Does the graph suggest that knowing a professional golfer's score for one round is much help in predicting his score for another round on the same course?

4.25 Happy states. Human happiness or well-being can be assessed either subjectively or objectively. Subjective assessment can be accomplished by listening to what people say. Objective assessment can be made from data related to well-being such as income, climate, availability of entertainment, housing prices, lack of traffic congestion, etc. Do subjective and objec-

FIGURE 4.7
Scatterplot of school GPA against IQ test scores for seventh-grade students, for Exercises 4.16 and 4.17.

FIGURE 4.8
Scatterplot of the scores in the first two rounds of the 2010 Masters Tournament, for Exercise 4.24.
system of health surveys. Lower scores indicate a greater degree of happiness. To objectively assess happiness, the investigators computed a mean well-being score (called the compensating-differentials score) for each state, based on objective measures that have been found to be related to happiness or well-being. The states were then ranked according to this score (Rank 1 being the happiest). Figure 4.9 is a scatterplot of mean BRFSS scores (response) against the rank based on the compensating-differentials scores (explanatory).

(a) Is there an overall positive association or an overall negative association between mean BRFSS score and rank based on the compensating-differentials method? (b) Does the overall association indicate agreement or disagreement between the mean subjective BRFSS score and the ranking based on objective data used in the compensating-differentials method? (c) Are there any outliers? If so, what are the BRFSS scores corresponding to these outliers?

4.26 Wine and cancer in women. Some studies have suggested that a nightly glass of wine may not only take the edge off a day but also improve health. Is wine good for your health? A study of nearly 1.3 million middle-aged British women examined wine consumption and the risk of breast cancer. The researchers were interested in how risk changed as wine consumption increased. Risk is based on breast cancer rates in drinkers relative to breast cancer rates in nondrinkers in the study, with higher values indicating greater risk. In particular, a value greater than 1 indicates a greater breast cancer rate than that of nondrinkers. Wine intake is the mean wine intake, in grams per day, of all women in the study who drank approximately the same amount of wine per week. Here are the data (for drinkers only):

<table>
<thead>
<tr>
<th>Wine intake x (grams per day)</th>
<th>2.5</th>
<th>8.5</th>
<th>15.5</th>
<th>26.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative risk y</td>
<td>1.00</td>
<td>1.08</td>
<td>1.15</td>
<td>1.22</td>
</tr>
</tbody>
</table>

(a) Make a scatterplot of these data. Based on the scatterplot, do you expect the correlation to be positive or negative? Near ±1 or not? (b) Find the correlation r between wine intake and relative risk. Do the data show that women who consume more wine tend to have higher relative risks of breast cancer?

4.27 Ebola and gorillas. The deadly Ebola virus is a threat to both people and gorillas in Central Africa. An outbreak in 2002 and 2003 killed 91 of the 95 gorillas in 7 home ranges in the Congo. To study the spread of the virus, measure “distance” by the number of home ranges separating a group of gorillas from the first group infected. Here are data on distance and time in number of days until deaths began in each later group:

<table>
<thead>
<tr>
<th>Distance</th>
<th>1</th>
<th>3</th>
<th>4</th>
<th>4</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>4</td>
<td>21</td>
<td>33</td>
<td>41</td>
<td>43</td>
<td>46</td>
</tr>
</tbody>
</table>

(a) Make a scatterplot. Which is the explanatory variable? What kind of pattern does your plot show? (b) Find the correlation r between distance and time. (c) If time in days were replaced by time in number of weeks until death began in each later group (fractions allowed so that 4 days becomes 4/7 weeks), would the correlation between distance and time change? Explain your answer.

4.28 Sparrowhawk colonies. One of nature’s patterns connects the percent of adult birds in a colony that return from the previous year and the number of new adults that join the colony. Here are data for 13 colonies of sparrowhawks:

- Morales Morales/Photolibrary
(a) Make a plot of percentage tip against the weather report on the bill (space the three weather reports equally on the horizontal axis). Which weather report appears to lead to the best tip?
(b) Does it make sense to speak of a positive or negative association between weather report and percentage tip? Why? Is correlation $r$ a helpful description of the relationship? Why?

4.34 Thinking about correlation. Exercise 4.26 presents data on wine intake and the relative risk of breast cancer in women.

(a) If wine intake is measured in ounces per day rather than grams per day, how would the correlation change? (There are 0.035 ounces in a gram.)
(b) How would $r$ change if all the relative risks were 0.25 less than the values given in the table? Does the correlation tell us that among women who drink, those who drink more wine tend to have a greater relative risk of cancer than women who don’t drink at all?
(c) If drinking an additional gram of wine each day raised the relative risk of breast cancer by exactly 0.01, what would be the correlation between wine intake and relative risk of breast cancer? (Hint: Draw a scatterplot for several values of wine intake.)

4.35 The effect of changing units. Changing the units of measurement can dramatically alter the appearance of a scatterplot. Return to the data on knee height and overall height in Exercise 4.23: ❄️ KNEEH2

<table>
<thead>
<tr>
<th>Knee height $x$</th>
<th>57.7</th>
<th>47.4</th>
<th>43.5</th>
<th>44.8</th>
<th>55.2</th>
<th>54.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height $y$</td>
<td>192.1</td>
<td>153.3</td>
<td>146.4</td>
<td>162.7</td>
<td>169.1</td>
<td>177.8</td>
</tr>
</tbody>
</table>

Both heights are measured in centimeters. A mad scientist decides to measure knee height in millimeters and height in meters. The same data in these units are

<table>
<thead>
<tr>
<th>Knee height $x$</th>
<th>577</th>
<th>474</th>
<th>435</th>
<th>448</th>
<th>552</th>
<th>546</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height $y$</td>
<td>1.921</td>
<td>1.533</td>
<td>1.464</td>
<td>1.627</td>
<td>1.691</td>
<td>1.778</td>
</tr>
</tbody>
</table>

(a) Make a plot with the $x$ axis extending from 0 to 600 and the $y$ axis from 0 to 250. Plot the original data on these axes. Then plot the new data using a different color or symbol. The two plots look very different.
(b) Nonetheless, the correlation is exactly the same for the two sets of measurements. Why do you know that this is true without doing any calculations? Find the two correlations to verify that they are the same.

4.36 Statistics for investing. Investment reports now often include correlations. Following a table of correlations among mutual funds, a report adds: “Two funds can have perfect correlation, yet different levels of risk. For example, Fund A and Fund B may be perfectly correlated, yet Fund A moves 20% whenever Fund B moves 10%.” Write a brief explanation, for someone who knows no statistics, of how this can happen. Include a sketch to illustrate your explanation.

4.37 Statistics for investing. A mutual funds company’s newsletter says, “A well-diversified portfolio includes assets with low correlations.” The newsletter includes a table of correlations between the returns on various classes of investments. For example, the correlation between municipal bonds and large-cap stocks is 0.50, and the correlation between municipal bonds and small-cap stocks is 0.21.

(a) Rachel invests heavily in municipal bonds. She wants to diversify by adding an investment whose returns do not closely follow the returns on her bonds. Should she choose large-cap stocks or small-cap stocks for this purpose? Explain your answer.
(b) If Rachel wants an investment that tends to increase when the return on her bonds drops, what kind of correlation should she look for?

4.38 Teaching and research. A college newspaper interviews a psychologist about student ratings of the teaching of faculty members. The psychologist says, “The evidence indicates that the correlation between the research productivity and teaching rating of faculty members is close to zero.” The paper reports this as “Professor McDaniel said that good researchers tend to be poor teachers, and vice versa.” Explain why the paper’s report is wrong. Write a statement in plain language (don’t use the word “correlation”) to explain the psychologist’s meaning.

4.39 Sloppy writing about correlation. Each of the following statements contains a blunder. Explain in each case what is wrong.

(a) “There is a high correlation between the sex of American workers and their income.”
(b) “We found a high correlation ($r = 1.09$) between students’ ratings of faculty teaching and ratings made by other faculty members.”
(c) “The correlation between height and weight of the subjects was $r = 0.63$ centimeter.”

4.40 Correlation is not resistant. Go to the Correlation and Regression applet. Click on the scatterplot to create a group of 10 points in the lower-left corner of the scatterplot with a strong straight-line pattern (correlation about 0.9).
CHECK YOUR SKILLS

5.18 Figure 5.9 is a scatterplot of school GPA against IQ test scores for 15 seventh-grade students. The line is the least-squares regression line for predicting school GPA from IQ score. If another child in this class has IQ score 110, you predict the school GPA to be close to
(a) 2. (b) 7.5. (c) 11.

**FIGURE 5.9**
Scatterplot of IQ test scores and school GPA for 15 seventh-grade students, for Exercises 5.18 and 5.19.

5.19 The slope of the line in Figure 5.9 is closest to
(a) -11. (b) 0.2. (c) 2.0.

5.20 The points on a scatterplot lie close to the line whose equation is \( y = 4 - 3x \). The slope of this line is
(a) 4. (b) 3. (c) -3.

5.21 Fred keeps his savings in his mattress. He began with $1000 from his mother and adds $100 each year. His total savings \( y \) after \( x \) years are given by the equation
(a) \( y = 1000 + 100x \). (b) \( y = 100 + 1000x \).
(c) \( y = 1000 + x \).

5.22 Smokers don’t live as long (on the average) as non-smokers, and heavy smokers don’t live as long as light smokers. You regress the age at death of a group of male smokers on the number of packs per day they smoked. The slope of your regression line
(a) will be greater than 0. (b) will be less than 0.
(c) Can’t tell without seeing the data.

5.23 An owner of a home in the Midwest installed solar panels to reduce heating costs. After installing the solar panels, he measured the amount of natural gas used \( y \) (in cubic feet) to heat the home and outside temperature \( x \) (in degree-days, where a day’s degree-days are the number of degrees its average temperature falls below 65°F) over a 23-month period. He then computed the least-squares regression line for predicting \( y \) from \( x \) and found it to be
\[
\hat{y} = 85 + 16x
\]
How much, on average, does gas used increase for each additional degree-day?
(a) 23 cubic feet (b) 85 cubic feet (c) 16 cubic feet

5.24 According to the regression line in Exercise 5.23, the predicted amount of gas used when the outside temperature is 20 degree-days is about
(a) 405 cubic feet. (b) 320 cubic feet.
(c) 105 cubic feet.

5.25 By looking at the equation of the least-squares regression line in Exercise 5.23, you can see that the correlation between amount of gas used and degree-days is
(a) greater than zero. (b) less than zero.
(c) Can’t tell without seeing the data.

5.26 The software used to compute the least-squares regression line in Exercise 5.23 says that \( r^2 = 0.98 \). This suggests that
(a) although degree-days and gas used are correlated, degree-days does not predict gas used very accurately.
(b) gas used increases by \( \sqrt{0.98} = 0.99 \) cubic feet for each additional degree-day.
(c) prediction of gas used from degree-days will be quite accurate.

5.27 Because elderly people may have difficulty standing to have their heights measured, a study looked at predicting overall height from knee to the height. Here are data (in centimeters) for six elderly men:

<table>
<thead>
<tr>
<th>Knee height x</th>
<th>57.7</th>
<th>47.4</th>
<th>43.5</th>
<th>44.8</th>
<th>55.2</th>
<th>54.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height y</td>
<td>192.1</td>
<td>153.3</td>
<td>146.4</td>
<td>162.7</td>
<td>169.1</td>
<td>177.8</td>
</tr>
</tbody>
</table>

Use your calculator or software: what is the equation of the least-squares regression line for predicting overall height from knee height?
(a) \( \hat{y} = 42.9 + 2.5x \) (b) \( \hat{y} = -3.4 + 0.3x \)
(c) \( \hat{y} = 2.5 + 42.9x \)
(a) What is the equation of the least-squares regression line for predicting beak heat loss, as a percent of total body heat loss from all sources, from temperature? Use the equation to predict beak heat loss, as a percent of total body heat loss from all sources, at a temperature of 25 degrees Celsius.

(b) What percent of the variation in beak heat loss is explained by the straight-line relationship with temperature?

(c) Use the information in Figure 5.11 to find the correlation $r$ between beak heat loss and temperature. How do you know whether the sign of $r$ is + or −?

5.32 Husbands and wives. The mean height of American women in their twenties is about 64.3 inches, and the standard deviation is about 3.9 inches. The mean height of men the same age is about 69.9 inches, with standard deviation about 3.1 inches. Suppose that the correlation between the heights of husbands and wives is about $r = 0.5$.

(a) What are the slope and intercept of the regression line of the husband's height on the wife's height in young couples?

(b) Draw a graph of this regression line for heights of wives between 56 and 72 inches. Predict the height of the husband of a woman who is 67 inches tall, and plot the wife's height and predicted husband's height on your graph.

(c) You don't expect this prediction for a single couple to be very accurate. Why not?

5.33 What's my grade? In Professor Krugman's economics course the correlation between the students' total scores prior to the final examination and their final-examination scores is $r = 0.5$. The pre-exam totals for all students in the course have mean 280 and standard deviation 40. The final-exam scores have mean 75 and standard deviation 8. Professor Krugman has lost Julie's final exam but knows that her total before the exam was 300. He decides to predict her final-exam score from her pre-exam total.

(a) What is the slope of the least-squares regression line of final-exam scores on pre-exam total scores in this course? What is the intercept?

(b) Use the regression line to predict Julie's final-exam score.

(c) Julie doesn't think this method accurately predicts how well she did on the final exam. Use $r$ to argue that her actual score could have been much higher (or much lower) than the predicted value.

5.34 Going to class. A study of class attendance and grades among first-year students at a state university showed that, in general, students who attended a higher percent of their classes earned higher grades. Class attendance explained 16% of the variation in grade index among the students.

What is the numerical value of the correlation between percent of classes attended and grade index?

5.35 Sisters and brothers. How strongly do physical characteristics of sisters and brothers correlate? Here are data on the heights (in inches) of 12 adult pairs:

<table>
<thead>
<tr>
<th>Brother</th>
<th>71</th>
<th>68</th>
<th>66</th>
<th>67</th>
<th>70</th>
<th>71</th>
<th>70</th>
<th>73</th>
<th>72</th>
<th>65</th>
<th>66</th>
<th>70</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sister</td>
<td>69</td>
<td>64</td>
<td>65</td>
<td>63</td>
<td>65</td>
<td>62</td>
<td>65</td>
<td>64</td>
<td>66</td>
<td>59</td>
<td>62</td>
<td>64</td>
</tr>
</tbody>
</table>

(a) Use your calculator or software to find the correlation and the equation of the least-squares line for predicting sister's height from brother's height. Make a scatterplot of the data and add the regression line to your plot.

(b) Damien is 70 inches tall. Predict the height of his sister Tonya. Based on the scatterplot and the correlation $r$, do you expect your prediction to be very accurate? Why?

5.36 Keeping water clean. Keeping water supplies clean requires regular measurement of levels of pollutants. The measurements are indirect—a typical analysis involves forming a dye by a chemical reaction with the dissolved pollutant, then passing light through the solution and measuring its "absorbence." To calibrate such measurements, the laboratory measures known standard solutions and uses regression to relate absorbence and pollutant concentration. This is usually done every day. Here is one series of data on the absorbence for different levels of nitrates. Nitrates are measured in milligrams per liter of water.

<table>
<thead>
<tr>
<th>Nitrates</th>
<th>50</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>400</th>
<th>800</th>
<th>1200</th>
<th>1600</th>
<th>2000</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absorbence</td>
<td>7.0</td>
<td>7.5</td>
<td>12.8</td>
<td>24.0</td>
<td>47.0</td>
<td>93.0</td>
<td>138.0</td>
<td>183.0</td>
<td>230.0</td>
<td>226.0</td>
</tr>
</tbody>
</table>

(a) Chemical theory says that these data should lie on a straight line. If the correlation is not at least 0.997, something went wrong and the calibration procedure is repeated. Plot the data and find the correlation. Must the calibration be done again?

(b) The calibration process sets nitrate level and measures absorbence. The linear relationship that results is used to estimate the nitrate level in water from a measurement of absorbence. What is the equation of the line used to estimate nitrate level? What does the slope of this line say about the relationship between nitrate level and absorbence? What is the estimated nitrate level in a water specimen with absorbence 40?

(c) Do you expect estimates of nitrate level from absorbence to be quite accurate? Why?

5.37 Sparrowhawk colonies. One of nature's patterns connects the percent of adult birds in a colony that return from the previous year and the number of new adults that join the colony. Here are data for 13 colonies of sparrowhawks:
You saw in Exercise 4.28 that there is a moderately strong linear relationship, with correlation \( r = -0.748 \).

(a) Find the least-squares regression line for predicting \( y \) from \( x \). Make a scatterplot and draw your line on the plot.

(b) Explain in words what the slope of the regression line tells us.

(c) An ecologist uses the line, based on 13 colonies, to predict how many new birds will join another colony, to which 60% of the adults from the previous year return. What is the prediction?

5.38 Our brains don’t like losses. Exercise 4.29 (page 117) describes an experiment that showed a linear relationship between how sensitive people are to monetary losses (“behavioral loss aversion”) and activity in one part of their brains (“neural loss aversion”).

(a) Make a scatterplot with neural loss aversion as \( x \) and behavioral loss aversion as \( y \). One point is a high outlier in both the \( x \) and \( y \) directions.

(b) Find the least-squares line for predicting \( y \) from \( x \), leaving out the outlier, and add the line to your plot.

(c) The outlier lies very close to your regression line. Looking at the plot, you now expect that adding the outlier will increase the correlation but will have little effect on the least-squares line. Explain why.

(d) Find the correlation and the equation of the least-squares line with and without the outlier. Your results verify the expectations from (c).

5.39 Always plot your data! Table 5.1 presents four sets of data prepared by the statistician Frank Anscombe to illustrate the dangers of calculating without first plotting the data.15

(a) Without making scatterplots, find the correlation and the least-squares regression line for all four data sets. What do you notice? Use the regression line to predict \( y \) for \( x = 10 \).

(b) Make a scatterplot for each of the data sets and add the regression line to each plot.

(c) In which of the four cases would you be willing to use the regression line to describe the dependence of \( y \) on \( x \)? Explain your answer in each case.

5.40 Managing diabetes. People with diabetes must manage their blood sugar levels carefully. They measure their fasting plasma glucose (FPG) several times a day with a glucose meter. Another measurement, made at regular medical

<table>
<thead>
<tr>
<th>TABLE 5.1</th>
<th>Four data sets for exploring correlation and regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x )</td>
<td>10  8  13  9  11  14  6  4  12  7  5</td>
</tr>
<tr>
<td>( y )</td>
<td>8.04 6.95 7.58 8.81 8.33 9.96 7.24 4.26 10.84 4.82 5.68</td>
</tr>
<tr>
<td>( x )</td>
<td>10  8  13  9  11  14  6  4  12  7  5</td>
</tr>
<tr>
<td>( x )</td>
<td>10  8  13  9  11  14  6  4  12  7  5</td>
</tr>
<tr>
<td>( y )</td>
<td>7.46 6.77 12.74 7.11 7.81 8.84 6.08 5.39 8.15 6.42 5.73</td>
</tr>
<tr>
<td>( x )</td>
<td>8  8  8  8  8  8  8  8  8  8  19</td>
</tr>
<tr>
<td>( y )</td>
<td>6.58 5.76 7.71 8.84 8.47 7.04 5.25 5.56 7.91 6.89 12.50</td>
</tr>
</tbody>
</table>