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—is 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 score of 1190 compares with those of the students who attend that university. If all you have is raw data—a list of the SAT I scores of the 22,000 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 their college guide, which says “the middle 50% of the scores were between 1170 and 1340, with half the scores above 1210 and half below.” Now you know that although your 1190 is in the bottom half of the scores, it is not far from the center value of 1210 and higher than the bottom quarter.

Notice that the summary of the scores gives you two different kinds of information: the **center** 1210 and the **spread** from 1170 to 1340, for the middle 50%. 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 the question “If you choose 3 people at random from a set of 10 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?” But generally you don’t need to do all this work! Using shape, center, and spread, it is possible to get an answer without doing a simulation. This remarkable fact first began to come to light in the late 1600s and 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 these facts.

**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 central or typical value
- choose and compute a useful measure of spread (variability)
- work with the normal, or bell-shaped, curve
Summaries simplify. In fact, summaries can sometimes oversimplify, which means that 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.)

Activity 2.1 introduces one of the most important common shapes 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 next activity, you’ll measure a tennis ball with a ruler, but the results you’ll get 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 its weekly weighing, they face a similar challenge because of variability.

**Activity 2.1 Measuring Diameters**

What you’ll need: a tennis ball and 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.
2. Using your method, make two measurements of the diameter of your tennis ball to the nearest millimeter.
3. Combine your data with that of the rest of the class to form a dot plot. Speculate first, though, 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 —??—.” (There is more than one reasonable set of choices.)
6. Discuss some possible reasons for the variability in the measurements. How could the variability be reduced? Can the variability be eliminated entirely? (We will return to these issues in Chapter 4.)

Distributions come in a variety of shapes, but four of the most common basic shapes are illustrated in the rest of this section.
Is there any reason to believe that more babies are born in one month than in another? Or should the number of births be fairly uniform across the year? Display 2.1 shows the U.S. births and deaths (in thousands) for 1997. Display 2.2 shows a plot of the birth data, along with a smooth approximation of the distribution.

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. You can use the smooth approximation as the basis for a short verbal summary: “The distribution of births is roughly uniform over the months January through December, with about 325,000 births per month.”

<table>
<thead>
<tr>
<th>Month</th>
<th>Births (in thousands)</th>
<th>Deaths (in thousands)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>305</td>
<td>218</td>
</tr>
<tr>
<td>2</td>
<td>289</td>
<td>191</td>
</tr>
<tr>
<td>3</td>
<td>313</td>
<td>198</td>
</tr>
<tr>
<td>4</td>
<td>342</td>
<td>189</td>
</tr>
<tr>
<td>5</td>
<td>311</td>
<td>195</td>
</tr>
<tr>
<td>6</td>
<td>324</td>
<td>182</td>
</tr>
<tr>
<td>7</td>
<td>345</td>
<td>192</td>
</tr>
<tr>
<td>8</td>
<td>341</td>
<td>178</td>
</tr>
<tr>
<td>9</td>
<td>353</td>
<td>176</td>
</tr>
<tr>
<td>10</td>
<td>329</td>
<td>193</td>
</tr>
<tr>
<td>11</td>
<td>304</td>
<td>189</td>
</tr>
<tr>
<td>12</td>
<td>324</td>
<td>192</td>
</tr>
</tbody>
</table>

Display 2.1  Births and deaths in the United States, 1997.
Source: Centers for Disease Control and Prevention.

Display 2.2  Births per month, 1997. An example of a (roughly) uniform distribution.
Computers and many calculators generate random numbers between 0 and 1 that have a uniform distribution. Display 2.3 shows a dot plot of 1000 random numbers generated by Minitab statistical software. The flat line across the top is a smoothed version of the plot. For this smooth approximation, the percentage of outcomes in any interval, such as \([0.2, 0.4]\), is given by the percentage of the total area that lies above the interval. Because 20% of the total area lies above the interval \([0.2, 0.4]\), the smooth approximation tells us that 20% of the random numbers fell between 0.2 and 0.4. (You’ll learn more about this kind of graph in the next section.)

Discussion: Uniform Distribution

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

D2. Think of scenarios that you would expect to give rise to 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

Practice

P1. Plot the number of deaths per month given in Display 2.1. Do they appear to be uniformly distributed over the months? Use your plot as the basis for a verbal summary of the way deaths are distributed over the months of the year.

P2. Display 2.3 shows 1000 numbers randomly selected from a uniform distribution on the interval \([0, 1]\). Now imagine a uniform distribution on \([0, 2]\).
   a. What value divides the plot in half, with half the numbers below that value, half above?
   b. What values divide the area into quarters?
c. What values enclose the middle 50% of the data?
d. What percentage of the values lie between 0.4 and 0.7?
e. What values enclose the middle 95% of the data?

**Normal Distribution**

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

Pennies minted in the United States are supposed to weigh 3.110 grams, but a tolerance of 0.130 grams is allowed in either direction. Display 2.4 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 examples match 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 shown in Display 2.5. 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, stretching infinitely far in both directions. On either side of the mode, at about 60% of the height of the highest point of the curve, are points of inflection, where the curve changes from concave down to concave up.

Display 2.4 Weights of pennies.


Display 2.5 A normal curve, showing the line of symmetry and points of inflection.
To estimate the center and spread for a normal distribution, start with the line of symmetry. The point where it cuts the $x$-axis is the mean (or average). This value is where the area under the curve would balance if you cut it out of cardboard and held a finger under it. For all normal distributions, the mode and mean are equal. To measure spread, estimate the horizontal distance from the line of symmetry to either point of inflection. This distance is called the standard deviation, or $SD$ for short.

**Example: Averages of Random Samples**

Display 2.6 shows the distribution of average ages computed from 1000 sets of 5 workers chosen at random from the 10 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.

**Solution**

The curve shown in the display has center at 46.5 and inflection points at 42.5 and 50.5. Thus, the estimated mean is 46.5, and the estimated standard deviation is 4. A typical random sample of 5 workers has an average age of 46.5, give or take 4 years or so.

It is difficult to locate inflection points, especially when curves are drawn by hand, so a more reliable way to estimate the standard deviation is to use areas. For a normal curve, roughly $\frac{2}{3}$ of the total area under the curve is between the vertical lines through the two inflection points. In other words, the interval that stretches for one standard deviation on either side of the mean accounts for roughly $\frac{2}{3}$ of the area. For the distribution in Display 2.6, roughly $\frac{2}{3}$ of the dots are in the interval 46.5 ± 4 or [42.5, 50.5].

Activity 2.1 and the last two examples together illustrate the three most common ways that 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 of random samples (average ages)

All three scenarios are quite common, which makes the normal distribution especially important in statistics.
Discussion: Normal Distribution

D3. Determine these summaries visually.
   a. Estimate the center and spread for the penny weight data in Display 2.4, and use your estimates to write a summary sentence.
   b. Estimate the mean and standard deviation for your class data from Activity 2.1.

Practice

P3. Sketch a normal distribution with mean 0 and standard deviation 1. This distribution is called a standard normal distribution.

P4. For each of the normal distributions in Display 2.7, 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.” Then check to see that this interval contains roughly $\frac{3}{4}$ of the total area under the curve.
   a. SAT verbal scores
   b. ACT scores
   c. heights of women attending college
   d. single-season batting averages for professional baseball players in the decade of the 1910s

Display 2.7  Four distributions that are approximately normal.
Skewed Distribution

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 of Display 2.8 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 pounds (this part of the distribution even looks approximately normal), but if someone shouts “Abnormal bear loose!” you had better run for cover because that unusual bear is likely to be big! The “unusualness” 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, like 0 for measurements and counts, or a maximum that values can’t go above, like 100 for percentages. For example, the distribution in Display 2.9 shows the grade-point averages of college students (mostly first-year students and sophomores) taking an introductory statistics course at the University of Florida during the spring of 1999. 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. The skew is to the left: An unusual GPA would be one that is low compared to most GPAs for students in the class.
For skewed distributions, the center and spread are not as clear-cut as they are for normal distributions. Because there is no line of symmetry, the idea of center is ambiguous. Moreover, because the left and right halves of a skewed distribution don’t match, “distance to the point of inflection” is ambiguous, too. To get around this problem, people often report the **quartiles**, three numbers that divide the values into fourths. This lets you describe a distribution as in the introduction to the chapter: “The middle 50% of the SAT scores were between 1170 and 1340, with half above 1210 and half below.”

To estimate these values from a dot plot, first draw a vertical line at the value that divides the dots into two halves. This value, called the **median**, is the measure of center. To measure spread, repeat the halving process with each half of the data: Draw a vertical line that cuts each half into two pieces with equal numbers of dots on either side. These values are the **lower quartile** and **upper quartile**. They enclose the middle 50% of the values.

**Example**

Divide the bears’ weights in Display 2.10 into four equal parts, and estimate the median and quartiles. Write a short summary of this distribution.

**Solution**

There are 143 dots in Display 2.10, 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. 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 pounds, with half above about 155 and half below.

**Discussion: Skewed Distribution**

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

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

D5. Make up a scenario (name the cases and variables) whose distribution you would expect to be skewed right because of a wall. What is responsible for the wall?

D6. Make up a scenario whose distribution you would expect to be skewed left because of a wall. What is responsible for the wall?

D7. Which would you expect to be the more common direction of skew, right or left? Why?

■ Practice

P5. Match each plot in Display 2.11 with its median and quartiles, that is, the set of values that divide the area into fourths.

a. 0.15, 0.50, 0.85
b. 0.50, 0.71, 0.87
c. 0.63, 0.79, 0.91
d. 0.35, 0.5, 0.65
e. 0.25, 0.50, 0.75

Display 2.11  Five distributions with different shapes.

P6. The U.S. Environmental Protection Agency’s National Priorities List Fact Book tells the number of hazardous waste sites for each of the U.S. states and territories. For 1992, the numbers ranged from 0 to 102, the middle 50% of the values were between 6 and 22, half were above 10, and half below. Sketch what the distribution might look like.


P7. Estimate the median and quartiles for the distribution of GPAs in Display 2.9. Then write a verbal summary of the same form as in the example.
**Bimodal Distribution**

Many distributions, including the normal, and many skewed distributions as well, have only one peak (unimodal), but some have two (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 2.12 shows the life expectancies for females from countries on two continents—Europe and Africa.

![Bimodal Distribution Image](image)

*Display 2.12*  Life expectancy of females by country on two continents.  

Europe and Africa are quite different in their socioeconomic conditions, and the life expectancies reflect those conditions. If you make separate plots for the two continents, the two peaks become essentially one peak in each plot, as shown in Display 2.13. And, yes, Europe is a mixture as well: east and west with means about 75 and 79, respectively.

![Display 2.13 Life expectancy of females in Africa and Europe](image)

*Display 2.13*  Life expectancy of females in Africa and Europe.

Although it makes sense to talk about the center of the distribution of life expectancies for Europe, or of those for Africa, notice that it doesn’t really make sense to talk about “the” center of the distribution for both continents together. Instead you could tell the locations of the two peaks. But finding the reason for the two modes and separating the cases into two distributions, tells 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 they 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 the lab. He saw a pattern in the results that you can observe in the plot of his data in Display 2.14.

Lord Rayleigh saw two clusters separated by a gap. (There is no formal definition of a gap or a cluster, so you will have to use your best judgment about them. For example, some people call a single outlier a cluster of one; others don’t. You could also 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 10 values to the left had all come from the chemically produced samples and the 9 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 like radon 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 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 is the horizontal distance from the center to the points of inflection, where the curvature changes. To estimate it, find the distance on either side of the mean that encloses about two-thirds of the cases.
• If your distribution is skewed, you can give the values (quartiles) that divide the distribution into fourths.
• If your distribution is bimodal, it isn’t useful to report a single center. 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 the chapter, you will study the various measures of center and spread in more detail and learn how to compute them.

### Exercises

**E1.** Sketch the shape you would expect each distribution to have.

a. Age of each person who died last year in the United States
b. Age of each person who got his or her 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 below as bimodal, skewed right, skewed left, approximately normal, or roughly uniform.

a. The incomes of the world’s 100 richest people
b. The birth rates of Africa and Europe
c. The heights of soccer players on the last U.S. Woman’s World Cup 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

**E3.** Sketch these distributions:

a. A uniform distribution that shows the sort of data you would get from rolling a fair die 6000 times
b. A roughly normal distribution with mean 15 and standard deviation 5
c. A distribution that is skewed left, with half its values above 20, half below, and that has the middle 50% of its values between 10 and 25
d. A distribution that is skewed right, with the middle 50% of its values between 100 and 1000 and with half the values above 200 and half below

d. A distribution that is skewed right, with the middle 50% of its values between 200 and 800 and with half the values above 500 and half below

**E4.** The plot in Display 2.15 shows the last digit of the social security numbers of the students in a statistics class. Describe this distribution.

**E5.** The dot plot in Display 2.16 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 this shape? Generate as many possibilities as you can.

E7. The distribution in Display 2.18 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 that shape?

Display 2.18  Strength of yarn.
Source: Data and Story Library at Carnegie Mellon University, http://lib.stat.cmu.edu/DASL.

E8. Although a uniform distribution gives a reasonably “smooth” approximation to the actual distribution of births over months (Display 2.2), 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.

E9. Draw a graph similar to that in Display 2.19 for the data on deaths in the United States in Display 2.1, and summarize what you find.

Display 2.20 gives the estimated number of viewers who watched each television program from start to finish. This week was special because it ended the season and featured the very last new episode of Seinfeld.

<table>
<thead>
<tr>
<th>Program</th>
<th>Network</th>
<th>Viewers (millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Seinfeld</td>
<td>NBC</td>
<td>76.26</td>
</tr>
<tr>
<td>2 Seinfeld Clips</td>
<td>NBC</td>
<td>58.53</td>
</tr>
<tr>
<td>3 ER</td>
<td>NBC</td>
<td>47.78</td>
</tr>
<tr>
<td>4 Touched by an Angel</td>
<td>CBS</td>
<td>20.47</td>
</tr>
<tr>
<td>5 The X-Files</td>
<td>FOX</td>
<td>18.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50 48 Hours</td>
<td>CBS</td>
<td>10.18</td>
</tr>
<tr>
<td>51 Dr. Quinn Medicine Woman</td>
<td>CBS</td>
<td>10.15</td>
</tr>
<tr>
<td>52 Beverly Hills, 90210</td>
<td>FOX</td>
<td>10.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>101 Malcolm and Eddie (Tue.)</td>
<td>UPN</td>
<td>2.32</td>
</tr>
</tbody>
</table>

Display 2.20 Nielsen estimates of television show viewers.

E10. The dot plot in Display 2.21 shows the distribution of the Nielsen ratings.

a. In the Nielsen data, what are the cases? Describe the variables.

b. Describe the basic shape of the distribution in Display 2.21. Note any outliers and any gaps or clusters in the distribution.

c. Find the median number of people who watched a prime-time television show. Is there a lot of spread (variability) in the numbers of viewers? The middle half of the ratings are between what two values?

d. What can you say about how the number of people watching the last episode of Seinfeld compared to the number who watch a typical television show?

e. The dot plot in Display 2.22 shows the Nielsen estimates of viewers for an ordinary week for which there was nothing special, such as the last Seinfeld episode. Compare the shape, center, and spread of this distribution with the one in Display 2.21.

Display 2.21 Number of viewers of television shows in millions, per Nielsen ratings.

Display 2.22 Dot plot of Nielsen ratings for a less unusual week.
E11. The dot plots in Display 2.23 can be used to compare the distributions of the ratings for the six networks.

a. Describe the basic shape of the distribution for each network. Note any outliers and any gaps or clusters in the distribution.

b. Compare the center and spread of the ratings for FOX and for NBC. For which of the six networks are the ratings centered highest? Lowest?

c. Which network has the most variability in its ratings? The least variability?

d. From looking at the plots, rank the six networks according to the popularity of their shows.

---

2.2 Graphical Displays for Distributions

Plots should present the essentials quickly and clearly.

As you saw in the last 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.

Pet cats typically live about 12 years, but some have been known to live for 28 years. Is that typical of domesticated predators? What about domesticated nonpredators, like cows and guinea pigs? Or wild mammals? The rhinoceros, a nonpredator, lives an average of 15 years, with a maximum of about 45 years. On the other hand, 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.
2.2 Graphical Displays for Distributions

<table>
<thead>
<tr>
<th>Mammal</th>
<th>Gestation Period (days)</th>
<th>Average Life Span (years)</th>
<th>Maximum Life Span (years)</th>
<th>Speed (mph)</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>Grizzly bear</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>14</td>
<td>70</td>
<td>1</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>

Display 2.24 Facts on mammals.

Cases and Variables, Quantitative and Categorical

Many of the examples in this section are based on the data about mammals in Display 2.24. For wild mammals, longevity is taken from records kept on mammals in captivity, and maximum longevity is the largest longevity on record. The column Wild is coded 1 if the mammal is wild and 0 if it is domestic. The column Predator is coded 1 if the mammal preys on other animals for food and 0 if it does not. The asterisks (*) mark missing values.

In Display 2.24, each row (each mammal) is a case. In general, the cases in a data set are the individual people, cities, mammals, or other items being studied. Measurements and other properties of the cases are organized into columns, one column for each variable. Thus, average longevity and speed are variables, and, for example, 30 mph is the value of the variable speed for the case grizzly bear. Speed is a quantitative variable because the speeds are numbers that can be compared in a meaningful way. Wild is a categorical variable, as is predator—although the values 0 and 1 are numbers, the numbers are actually substitutes for the categories “no” and “yes.”

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 reading 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.

As you saw in Section 2.1, a dot plot shows shape, center, and spread. They 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
Discussion: More About Dot Plots

D8. Classify each variable in Display 2.24 as quantitative or categorical.
D9. Consider the mammals’ speeds in Display 2.24.
   a. Count the number of mammals that have speeds ending in a 0 or a 5.
   b. How many would you expect to end in a 0 or a 5 just by chance?
   c. What are some possible explanations for the fact that your answers in
      parts a and b are so different?

Practice
P8. In the listing of the Westvaco data in Chapter 1 on page 5, which
    variables are quantitative? Which are categorical?
P9. Decide on a reasonable scale, and make a dot plot of the gestation periods
    of the mammals listed in Display 2.24. Describe the shape, center, and
    spread from this dot plot. 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

A dot plot shows individual cases as dots. A **histogram** shows groups of cases
as rectangles or bars. In fact, you can think of a histogram as a dot plot with
bars drawn around the dots and the dots erased. This makes the height of the
bar 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
(called **frequency** or count) that are represented by each bar. For example,
four mammals have speeds of 30 to 35 miles per hour.

![Histogram of mammal speeds.](display2.26)

Most statistical software places a value that falls at the dividing line between
two bars into the bar on the right. For example, in Display 2.26, the bar going
from 30 to 35 would contain values such that $30 \leq speed < 35$. 
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 and shows a more symmetric, bell-shaped distribution. Now there appears to be one peak rather than two. If there are few values in the data set, it is difficult to identify peaks. In this situation, it is better to use a plot that identifies individual values, like a dot plot or a stemplot.

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 version that shows important features of the data.

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 draw the plots for you

A histogram shows frequencies on the vertical axis. To make 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**

Display 2.28 shows the relative frequency distribution of life expectancies for 250 countries around the world. What proportion of the countries have life expectancies of 64 years or more?
2.2 Graphical Displays for Distributions

Solution

Locate the interval of values of 64 or more on the x-axis. What proportion of the total area is taken up by the bars over that interval? A rough visual estimate is about \( \frac{2}{3} \) of the area: Roughly \( \frac{2}{3} \) of the countries have life expectancies of at least 64 years. Now suppose you want a more precise estimate. The proportion of countries with life expectancies of 64 years or greater is the sum of the heights of the four bars of the histogram to the right of 64, or about 0.13 + 0.22 + 0.19 + 0.13 = 0.67.

Discussion: Histograms

D10. Describe the center and spread of the distribution of mammal speeds based first on the histogram in Display 2.26, then based on the histogram in Display 2.27. How much difference does the bar width make for this data set?

D11. 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 make all histograms with very narrow bars?

D12. Does using relative frequencies change the shape of a histogram? What information is lost or gained when presenting a relative frequency histogram rather than a frequency histogram?

Practice

P10. Using a calculator or computer, make histograms of the average longevities and the maximum longevities of the mammals. Describe how the distributions differ in terms of shape, center, and spread. Why do these differences occur?

P11. Convert your histograms of the average longevities and the maximum longevities of the mammals to relative frequency histograms. Do the shapes of the histograms change?
P12. In the histogram for life expectancies (Display 2.28), which will be larger, the mean (balance point) or the median (value that divides the area into a right half and a left half)? Explain your reasoning.

Stemplots

Both the dot plot and the histogram show the shape, center, and spread of a distribution of data, but neither retains the exact values. The plot in Display 2.29 shows the key features of the distribution and preserves all of the original numbers. It is a stem-and-leaf plot or stemplot of the mammal speeds.

Display 2.29  Stemplot of mammal speeds.

A stemplot shows cases as digits.

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 of 39 is printed in bold. If you turn your book 90° counterclockwise, you see what looks something like a dot plot or histogram, and you can see the shape, center, and spread of the distribution, just as you can from those plots.

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.

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 Display 2.13, for example). 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!

<table>
<thead>
<tr>
<th>Predator</th>
<th>Nonpredator</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 1</td>
<td>1 2</td>
</tr>
<tr>
<td>2 0</td>
<td>5</td>
</tr>
<tr>
<td>0 0 3 0 2</td>
<td>4 0 0 0</td>
</tr>
<tr>
<td>0 5</td>
<td>5 8</td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>0 7</td>
<td></td>
</tr>
</tbody>
</table>

3 9 represents 39 miles per hour

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 either be 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 miles per hour could either be truncated to 32 miles per hour or rounded to 33 miles per hour.

As with the other plots, the rules for making stemplots are flexible. Do what seems to work best to help your reader see 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 2.31, 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 2 cases on that stem. If you skip down three lines, the 4 on the left means that there are a total of 4 cases on the first 4 stems.
Stem-and-leaf of Speeds  \( N = 18 \)
Leaf Unit = 1.0  \( \text{N* = 21} \)

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaves</th>
</tr>
</thead>
<tbody>
<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>

Display 2.32  Stem-and-leaf plot of mammal speeds made by statistical software.

When are stem-and-leaf plots most useful?

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

D13. Describe the shape, center, and spread of the distribution of mammal speeds from the stemplot in Display 2.30 or Display 2.32. Compare your answer to that of D10.

D14. What information is given by the numbers in the leftmost column of the bottom half of the plot in Display 2.32?

D15. Discuss how you might construct a stemplot of the data on gestation periods for the mammals given in Display 2.24. Note that some of these values are three-digit numbers, so you will have to decide on a rule for stems and leaves.

Practice

P13. Make a back-to-back stemplot of the average longevities and maximum longevities from Display 2.24. Compare the two distributions.

P14. Examine Display 2.31 and describe how the speeds of predators and nonpredators seem to differ in terms of shape, center, and spread. Explain why you should expect these differences.
Bar graphs show frequencies for categorical data as heights of bars.

### Activity 2.2  Do Units of Measurement Affect Your Estimates?

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

1. Your instructor will randomly split the class 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 room in meters. Do this by looking at the length of the room; no pacing the length of the room allowed.
3. Find an appropriate and meaningful way to plot the two data sets so that you can compare them.
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 in meters?

### Bar Graphs for Categorical Data

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

The bar graph in Display 2.33 shows the frequency of mammals in the table that fall into the categories of wild and domestic. (Note that the bars are separated so that there is no suggestion that the variable can take on the value of, say, 1.5.)
Display 2.34 shows the proportion of the female labor force aged 25 and older in the United States that falls into various educational categories. The coding used in the plot 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 variable on the horizontal axis reflects the amount of formal education received. Even though it is labeled with numerical values here, attained education, as defined above, is best thought of as a categorical variable rather than a measurement. This bar graph, then, shows the relative frequencies for a categorical variable.

Discussion: Bar Graphs

D16. In the bar graph of Display 2.33, would it matter if the order of the bars were reversed? In the bar graph of Display 2.34, would it matter if the order of the first two bars in the graph were reversed? Comment on how we might define two different types of categorical variables.

D17. Examine the grouped bar graph in Display 2.35.
Display 2.35  Bar graph of frequency of wild and domestic mammals by predator status.

a. Describe what the height of each bar represents.

b. How can you tell from this bar graph whether a predator from our list is more likely to be wild or domestic?

c. How can you tell from this bar graph whether a nonpredator or a predator is more likely to be wild?

Practice

P15. Display 2.36 for 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 for female education compare to the distribution for male education? Why is it better to look at relative frequency bar graphs rather than frequency bar graphs to make this comparison?

Display 2.36  The male labor force 25 years and older by educational attainment.

P16. From the data in Display 2.23, make a bar graph showing the number of prime-time shows for each network.
Summary 2.2: Graphical Displays of Data

When a variable is quantitative, you can use dot plots, stemplots (or stem-and-leaf plots), and histograms to display the distribution of values. From each, you can see shape, center, and spread. 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 show approximations to the data values.
- Histograms show only intervals of values, losing the actual data values, and are most appropriate for large data sets.

A bar graph shows the distribution of a categorical variable.

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

- Where did this set of data come from?
- What are the cases and the variables?
- What is 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?

Exercises

E12. Suppose you collect this information for each student in your class: age, hair color, number of siblings, gender, miles he or she lives from school. What are the cases? What are the variables? Classify each variable as quantitative or categorical.

E13. The dot plot in Display 2.37 shows the distribution of the ages of the pennies in a sample collected by a statistics class.

a. Where did this set of data come from? What are the cases and the variables?

b. What are the shape, center, and spread of this distribution?

c. Does the distribution have any unusual characteristics? What are possible interpretations or explanations of the patterns you see in the distribution? That is, why does the distribution have the shape it does?

Display 2.37 Age of pennies. Each dot represents 4 points.

E14. How do you expect the distributions of average life expectancies to compare for wild and domesticated mammals?

a. Write your prediction in a sentence or two. Cover shape, center, and spread.
b. Use the data in Display 2.24 to make a back-to-back stemplot to compare average life expectancies.

c. Write a short summary comparing the two distributions.

E15. The graphs in Display 2.38 below appeared in a story on the “changing course of fast food.” What kinds of graphs are these? Study the graphs, and then write a story that might have been in the paper.

E16. Using your knowledge of the variables and what you think the shape of the distribution might look like, match each of the variables in the list below with the appropriate histogram in Display 2.39.

   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 2000 Summer Olympics
   IV. Weights of grown chickens in a barnyard

   Display 2.39 Four histograms with different shapes.

   A.  
   B.  
   C.  
   D.  

E17. Using the technology available to you, make histograms of the average longevity and maximum longevity data (Display 2.24) using bar widths of 4, 8, and 16 years. Comment on the main features of the shapes of these plots, and determine which bar width appears to display these features best.

---

**Number of Fast-Food Restaurants in the United States**

<table>
<thead>
<tr>
<th>Fast Food Chain</th>
<th>1992</th>
<th>1996</th>
</tr>
</thead>
<tbody>
<tr>
<td>McDonald’s</td>
<td>8,959</td>
<td>12,094</td>
</tr>
<tr>
<td>Burger King</td>
<td>5,705</td>
<td>7,057</td>
</tr>
<tr>
<td>Pizza Hut</td>
<td>7,609</td>
<td>8,701</td>
</tr>
<tr>
<td>Taco Bell</td>
<td>4,078</td>
<td>6,645</td>
</tr>
<tr>
<td>Wendy’s</td>
<td>3,607</td>
<td>4,369</td>
</tr>
</tbody>
</table>

**Change in Average Revenue per U.S. Restaurant Open at Least One Year**

<table>
<thead>
<tr>
<th>Chain</th>
<th>'93</th>
<th>'94</th>
<th>'95</th>
<th>'96</th>
</tr>
</thead>
<tbody>
<tr>
<td>McDonald’s</td>
<td>2.0%</td>
<td>-7.4%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burger King</td>
<td>-1.0%</td>
<td>2.7%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pizza Hut</td>
<td>6.4%</td>
<td>-7.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taco Bell</td>
<td></td>
<td></td>
<td>5.6%</td>
<td>-4.2%</td>
</tr>
<tr>
<td>Wendy’s</td>
<td>7.2%</td>
<td>1.2%</td>
<td>4%</td>
<td>-4%</td>
</tr>
</tbody>
</table>

*Display 2.38 Fast food restaurants.*
*Source: USA Today, June 6, 1997.*
E18. The histogram in Display 2.40 shows the distribution of average ages for 1000 random samples of size 3 chosen from the set of 10 hourly workers involved in the second round of layoffs at Westvaco.

a. Estimate the mean and standard deviation.

b. Very roughly, what percentage of the 1000 averages would you estimate are within one standard deviation of the mean? Within two standard deviations? Three standard deviations?

c. For this set of 1000 repetitions, about how many samples had an average age of 58 or more? What percentage of 1000 is this?


a. 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? Within two standard deviations? Three standard deviations?

c. For SAT I verbal scores, the shape was similar, but the mean was 9 points lower and the standard deviation was 2 points smaller. Draw a smooth curve to show the distribution of SAT I verbal scores.

E20. Display 2.42 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. Estimate the mean and standard deviation.
c. Estimate the proportion of men aged 18 to 24 who are 74 inches tall or less.

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

e. Explain why, in the histogram of Display 2.42, you can find proportions either by adding the heights of the bars or by adding the areas of the bars. Is this true of every histogram?

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

E21. The plots in Display 2.43 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.

E22. Look through newspapers and magazines to find an example of a graph that is either misleading or difficult to interpret. Redraw the graph to make it clear.

Display 2.43 Population pyramids for the United States and for Mexico for 2000.


2.3 Measures of Center and Spread

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

Measures of Center

The two most commonly used measures of center are the mean and the median.
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 2.44.) If a distribution is approximately normal, it balances at the line of symmetry, so the mean is on the horizontal axis directly below the highest point of the bell curve.

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

(The symbol \( \sum \), for sum, means to add up all of the values of \( x \).)

Display 2.44  The mean is the balance point of a distribution.

The median is the halfway point. The median is the value that divides the data into halves as shown in Display 2.45. To find it, list all of the values in order, and select the middle one, or the average of the two middle ones. If there are \( n \) values, you can find the median at, or surrounding, position \( \frac{n+1}{2} \).

Display 2.45  The median divides the distribution into two equal areas.
Example

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* (* means laid off in Round 2). The two dot plots in Display 2.46 show the distributions before and after the second round. What was the effect of Round 2 on the mean age? On the median age?

Solution

Means:

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

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

The layoffs reduced the mean age by 5 years.

Medians:

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

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

The layoffs reduced the median age by 13.5 years.

Discussion: Measures of Center

D18. Find the mean and median for each ordered list, and contrast their behavior.

a. 1 2 3  b. 1 2 6
   c. 1 2 9  d. 1 2 297

D19. As you saw in D18, typically the mean is more affected 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, the means are 43.5 mph for predators and 31.5 for nonpredators. The medians are 40.5 and 33.5. What is it about the distributions that causes the means to be farther apart than the medians?

c. What is it about the shapes of the plots in Display 2.46 that explains why the means change so much less than the medians?

Practice

P17. Find the mean and median of these ordered lists.

a. 1 2 3 4   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

P18. Five 3rd graders, all about 4 feet tall, are standing together when their teacher, who is 6 feet tall, joins the group. What happens to the mean height? The median height?

P19. The stemplots in Display 2.47 show the life expectancies (in years) for the population in the countries of Africa and Europe. The means are 53.6 years for Africa and 73.6 years for Europe.

a. Find the median of each data set.

b. Is the mean or the median smaller for each distribution? Why is this so?

Display 2.47  Life expectancies in Africa and Europe.

2.3 Measures of Center and Spread

Measuring Spread Around the Median: Quartiles and IQR

If you locate the center 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 fourths. The distance between the upper and lower quartiles, called the **interquartile range**, or **IQR**, is a measure of spread:

\[
IQR = Q_3 - Q_1
\]

The next example illustrates the value of the IQR. San Francisco, California, and Springfield, Missouri, have about the same average temperature across the year, a little above 55 degrees Fahrenheit. In San Francisco, half the months of the year have their normal temperatures above 56.5°F, half below. For Springfield, half the months have their normal temperatures above 57°F, half below. If you judge by these medians, the difference hardly matters. But if you visit San Francisco, you had better take a jacket, no matter what month you go. If you visit Springfield, however, 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 the normal monthly temperatures lie in a narrow 9-degree interval between 52.5°F and 61.5°F, whereas in Springfield, the middle 50% of the normal monthly temperatures range widely, over a 31-degree interval, from 40.5°F to 71.5°F. In short, the IQR is 9 degrees for San Francisco, 31 degrees for Springfield.

Finding the Quartiles

If you have an even number of cases, finding the quartiles is straightforward: Order your observations, divide them into a lower and upper half, then divide each half in half. If you have an odd number of cases, the idea is still 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, and you may get a slightly different value from some computer programs, but in this book the rule is to **omit the middle value when you form the two halves**.

**Example**

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

**Solution**

**Before**: There are 10 ages: 25, 33, 35, 38, 48, 55, 55, 55, 56, 64. Because \(n\) is even, the median is halfway between the two middle values. 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 \(Q_1 = 35\). 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 \(Q_3 = 55\).
Chapter 2: Exploring Distributions

After: After the three workers are laid off, there are 7 ages: 25, 33, 35, 38, 48, 55, 56. Because \( n \) is odd, the median is the middle value, or 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 \( Q_1 = 33 \). The upper half of the data is the set of the three values to the right of position 4, and the median of these is again the second value, so \( Q_3 = 55 \).

Discussion: Finding the Quartiles

D20. Here are the medians and quartiles for the speeds of the domestic and wild mammals:

<table>
<thead>
<tr>
<th></th>
<th>( Q_1 )</th>
<th>Median</th>
<th>( Q_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domestic</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 to verify these numbers, and then use them to summarize and compare the two distributions.

b. Why would the speeds of domestic mammals be less spread out than the speeds of wild mammals?

D21. The following quote comes from the mystery *The List of Adrian Messenger* by Philip MacDonald (Garden City, NY: Doubleday, 1959, page 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.

Practice

P20. Find the quartiles and \( IQRs \) 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
P21. Display 2.48 shows a back-to-back stemplot for the average life spans of predators and nonpredators.

<table>
<thead>
<tr>
<th>Predators</th>
<th>Nonpredators</th>
</tr>
</thead>
<tbody>
<tr>
<td>103</td>
<td>34</td>
</tr>
<tr>
<td>75</td>
<td>56788</td>
</tr>
<tr>
<td>22222</td>
<td>1000222</td>
</tr>
<tr>
<td>65</td>
<td>555555</td>
</tr>
<tr>
<td>2</td>
<td>0000</td>
</tr>
<tr>
<td>5·3</td>
<td>5</td>
</tr>
<tr>
<td>41</td>
<td></td>
</tr>
</tbody>
</table>

1|5 stands for 15 years

Display 2.48 Average life spans of predators and nonpredators.

a. Use the plot to find the medians and quartiles for each group of mammals.

b. Write a pair of sentences summarizing and comparing the two distributions.

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 in the set of data
- **Lower or first quartile, \(Q_1\)**: The median of the lower half of the values
- **Median**: The value that divides the data into halves
- **Upper or third quartile, \(Q_3\)**: The median of the upper half of the values
- **Maximum**: The largest value in the set of data

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

Display 2.49 shows the five-number summary for the speeds of the mammals listed in Display 2.24.

```
1 1 2 3 0
2 5 0 0 0 2 5 9
3 0 0 0 2 5 8
4 0 0 0 2 5 8
5 0 0
6 0
```

Display 2.49 Five-number summary for the mammal speeds.
A boxplot is sometimes referred to as a box and whiskers plot.

Display 2.50 is a boxplot for the mammal speeds. A boxplot is a graphical display of the five-number summary. The “box” extends from $Q_1$ to $Q_3$, with a line across it at the median. The “whiskers” run from the quartiles to the most extreme values.

The maximum speed of 70 mph for the cheetah is 20 mph from the next fastest mammal (the lion) and 28 mph from the nearest quartile. It is handy to have a version of the boxplot that shows isolated cases—outliers—like the cheetah. Informally, outliers are any values that stand apart from the rest, but you can use this rule to identify them:

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

Note that “more than 1.5 times the IQR from the nearest quartile” is another way of saying “either greater than $Q_3 + 1.5 \cdot IQR$, or less than $Q_1 – 1.5 \cdot IQR$.”

**Example**

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

**Solution**

From Display 2.49, $Q_1 = 30$ and $Q_3 = 42$, so the $IQR = 42 – 30 = 12$, and $1.5 \cdot IQR = 18$.

At the low end:

$Q_1 – 1.5 \cdot IQR = 30 – 18 = 12$

The pig, at 11 mph, is an outlier.

The squirrel, at 12 mph, is the smallest non-outlier.

At the high end:

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

The cheetah, at 70 mph, is an outlier.

The lion, at 50 mph, is the largest non-outlier.
A modified boxplot (shown in Display 2.51) 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.

Boxplots are particularly useful for comparing several distributions.

**Example**
Display 2.52 shows side-by-side modified boxplots of average longevity for wild and domestic mammals. Compare the two distributions.

**Solution**
The boxplot for domestic animals has no median line. So many domestic animals had an average longevity of 12 years that it is both the median and the upper quartile. Keeping that in mind, these plots show that, typically, species of domestic mammals have median average life spans of about 12 years, with about half of these average life spans falling between 8 and 12 years. The average life spans for wild mammals center at about the same place, but the wild mammal averages have more variability. The unusual average life spans are on the high side; two large mammals have average life spans of more than 30 years.

Boxplots are useful when you are plotting a single quantitative variable and
- you want to compare the shape, center, and spreads of two or more distributions
- your distribution has so many values that it would take too long, or use too much space, to show them individually in a stemplot
- 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 clearly indicated
Discussion: Five-Number Summaries, Outliers, and Boxplots

D22. Does the five-number summary give the position of the quartiles or the value of the quartiles, or is there any difference? What is another name for the second quartile?

D23. Test your ability to interpret boxplots with 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 from a boxplot without the five-number summary? How can you estimate the range?
   d. Contrast the information you can learn from a boxplot with that from a histogram. List the advantages and the disadvantages of each.

Practice

P22. Display 2.53 shows a boxplot of the Nielsen ratings from Display 2.20 and Display 2.21 of Section 2.1.

Display 2.53  Modified boxplot of Nielsen ratings.

a. Which three shows are the outliers?
   b. Which show is at the top of the upper whisker (the largest non-outlier)?
   c. Without looking back, sketch a histogram that could result in this boxplot.

P23. Use the medians and quartiles given in D20 and the data in Display 2.24 to construct side-by-side boxplots for the speeds of wild and domestic mammals. (Don’t show outliers in these plots.)

P24. The stemplot of average mammal life spans appears in Display 2.54.

Display 2.54  Average life span (in years) for 38 mammals.
a. Use it to find the five-number summary.
b. Find the IQR.
c. Compute $Q_1 - 1.5 \cdot IQR$. Identify any outliers (give the animal name and life span) at the low end.
d. Now identify an outlier at the high end and the largest non-outlier.

P25. Use your answers in P24 to draw a modified boxplot.

P26. Is it possible for a boxplot to be missing a whisker? If so, give an example. If not, explain why not.

**Percentiles and Cumulative Frequency Plots**

The first quartile, $Q_1$, of a distribution is the 25th percentile—the value that separates the lowest 25% of the data from the rest. The median is the 50th percentile, and $Q_3$ is the 75th percentile. In the same way, you can define other percentiles. The 10th percentile, for example, is the value that separates the bottom 10% of values in a distribution from the rest.

For large data sets, you may see data listed in a table or plotted in a graph like the SAT I verbal scores in Display 2.55. This plot is sometimes called a **cumulative percentage plot** or a **cumulative relative frequency plot**. The table shows that, for example, 30% of the students received a score of 450 or lower. About 14% 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>98</td>
</tr>
<tr>
<td>700</td>
<td>95</td>
</tr>
<tr>
<td>650</td>
<td>89</td>
</tr>
<tr>
<td>600</td>
<td>79</td>
</tr>
<tr>
<td>550</td>
<td>65</td>
</tr>
<tr>
<td>500</td>
<td>47</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Score</th>
<th>Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>450</td>
<td>30</td>
</tr>
<tr>
<td>400</td>
<td>16</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>

**Display 2.55** Cumulative relative frequency plot of SAT I verbal scores and percentiles, 1999–2000.


**Discussion: Percentiles and Cumulative Frequency Plots**

D24. Refer to Display 2.55.

a. Use the plot to estimate the percentile for an SAT I verbal score of 425.
b. What two values enclose the middle 90% of the SAT scores? The middle 95%?
c. Use the table to estimate the score that falls at the 40th percentile.
D25. What fraction of the cases lie between the 5th and 95th percentiles of a distribution? What percentiles enclose the middle 95% of the cases in a distribution?

Practice

P27. Estimate the quartiles and the median of the SAT I verbal scores in Display 2.55, and use those values to draw a boxplot for the distribution. What is the value of the IQR?

Measuring Spread Around the Mean: The Standard Deviation

There are various ways you can measure the spread of a distribution around its mean. The next activity will give you a chance to create a measure of your own.

Activity 2.3 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 the measurements aren’t all the same.
5. How far is your hand span from that of the mean of your group? How far from the mean are the hand spans of the others in your group?
6. Make a second plot, this time a dot 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 add to zero.

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

What is a typical deviation? As you saw in the activity, there are various ways to say what you mean by “typical,” but one measure, the standard deviation, abbreviated \( SD \), or \( s \), offers an important advantage you don’t get with other measures. There is a simple relationship between the standard deviation of a list of values and the standard deviation of the averages you get when you repeatedly choose random samples from the list. This reason for using the standard deviation depends on things you won’t learn about until Chapter 5. But you can get a preview of the basic idea if you turn back to Display 1.8, the simulation of the process of randomly choosing workers to lay off from Westvaco. If you’d had to do all those simulations by hand, you’d have been busy for quite a while, but there’s a shortcut. Unlike other measures of spread, you can compute the value of the standard deviation for the distribution of all those sample averages without doing any simulations. You only need to know two things: the number of workers you were choosing in each random sample and the standard deviation for the set of 10 workers you were choosing from. This remarkable property makes the standard deviation the most useful measure of spread for working with random samples.

To get these advantages, you have to work with squared deviations \( (x - \bar{x})^2 \). To compute the standard deviation, you first square the deviations, then take the average of those squares, and then take the square root.

Two versions of the standard deviation formula are used. One divides by the sample size \( n \) to get the average of the squared deviations; the other divides by \( n - 1 \). Your calculator probably computes both of these. (On some calculators, the two versions are labeled \( \sigma_n \) and \( \sigma_{n-1} \).) Dividing by \( n - 1 \) gives a slightly larger value for the standard deviation, and the larger value works better in statistical inference. If the choice makes much difference in the value of the standard deviation, however, your sample is probably too small for the standard deviation to be of much practical use anyway. For now, even though dividing by \( n \) may seem more natural, use \( n - 1 \) instead. We will come back to this in Chapter 5.

\[
\text{Formula for the Standard Deviation, } s \\
\sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}
\]

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

Compute the standard deviation for the average longevity of domesticated mammals from Display 2.24.

Solution

The table in Display 2.56 is a good way to organize the steps. First find the mean longevity \( \bar{x} \), 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 2.56  Computing the standard deviation.

To get the standard deviation, sum up the squared deviations, divide the sum by \( n - 1 \), and finally, take the square root:

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

Discussion: The Standard Deviation

D26. Does 4.67 years seem like a typical distance from the mean of 11 years for the average life spans in the example?

D27. 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?

D28. When you divide by \( n - 1 \) rather than by \( n \), what effect does it have on the standard deviation?

D29. 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?
Practice

P28. Verify that the sum of the deviations from the mean is 0 for the set 1, 2, 4, 6, 9. Find the standard deviation.

P29. Without computing, match each list of numbers on the left with its standard deviation in the right column. Check any answers you aren’t sure of by computing.

<table>
<thead>
<tr>
<th>List</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. 1 1 1 1</td>
<td>i. 0</td>
</tr>
<tr>
<td>b. 1 2 2</td>
<td>ii. 0.058</td>
</tr>
<tr>
<td>c. 1 2 3 4 5</td>
<td>iii. 0.577</td>
</tr>
<tr>
<td>d. 10 20 20</td>
<td>iv. 1.581</td>
</tr>
<tr>
<td>e. 0.1 0.2 0.2</td>
<td>v. 3.162</td>
</tr>
<tr>
<td>f. 0 2 4 6 8</td>
<td>vi. 3.606</td>
</tr>
<tr>
<td>g. 0 0 0 5 6 6 8 8</td>
<td>vii. 5.774</td>
</tr>
</tbody>
</table>

Properties of the Summary Statistics

Which summary statistics should you use to describe a distribution? Mean and standard deviation? Median and quartiles? Something else? The right choice depends on the shape of your distribution, so you should always start with a plot. For normal-shaped distributions, the mean and standard deviation are nearly always the most suitable summaries. For skewed distributions, the median and quartiles are often the most useful summaries, 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 choices 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. Later, when you study statistical inference, you’ll find that the standard deviation is the most useful measure of spread. This is because, as you saw in E18, the distribution of the sample means is approximately normal with a standard deviation that is easily estimated.

Choosing the right summaries is something you will get better at as you build your intuition about the properties of the summary statistics and how they behave in various situations.

Discussion: Which Summary Statistic?

D30. 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?

D31. When the “average” income of a community’s residents is given, that number is usually the median. Why do you think that is the case?

D32. Which summary statistics would be most useful in the following situations?

a. You are designing airline seats and want them to be wide enough for most people.
b. You are looking for the best buy on a specific type of calculator.
c. You would like to get a job when you start college but are unsure of how many hours you will need for study time.

**Practice**

P30. A community near Los Angeles has 9751 households with a median house price of $320,000 and an average price of $392,059. Why is the mean larger than the median? The property tax rate is about 1.15%. What is the total amount of taxes that will be assessed on these houses? What is the average amount per house?

P31. A story in the *Los Angeles Times* (July 30, 1998, page W14) reported that the median age of a car in 1997 was 8.1 years, the oldest ever. The medians were 6.5 years in 1990 and 4.9 years in 1970.

  a. Why were medians used in this story?
  
  b. What reasons might there be for the increase in median age of cars?

**The Effects of Recentering and Rescaling**

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

The lowest temperature on record for Washington, D.C., is –15°F. How does that compare with the lowest recorded temperatures for cities of other countries? Display 2.57 gives data for the few cities whose record temperatures turn out to be whole numbers in both the Fahrenheit and Celsius scales.

<table>
<thead>
<tr>
<th>City</th>
<th>Country</th>
<th>Temperature (°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>São Paulo</td>
<td>Brazil</td>
<td>32</td>
</tr>
<tr>
<td>Warsaw</td>
<td>Poland</td>
<td>–22</td>
</tr>
</tbody>
</table>

*Display 2.57* Record low temperatures for seven cities.


The dot plot in Display 2.58 shows that the temperatures are centered at about 32 with an outlier at –22. The spread and shape are hard to determine with only seven values.
What happens to the shape and spread of this distribution if you convert each temperature to number of degrees above or below freezing, 32°F? To find out, subtract 32 from each value, and plot the new values. Display 2.59 shows that the center of the dot plot is now at 0 rather than 32 but that 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 °C? The Celsius scale measures temperature using 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 2.59 by 1.8, and plot the new values. Display 2.60 shows that the center of the new dot plot is still at 0 and the shape is the same but the spread has decreased by a factor of 1.8.

Multiplying or dividing each value in a set of data by a positive constant doesn’t change the basic shape of the distribution. The mean and the spread both are multiplied by that number. Thus, this transformation amounts to a rescaling of the distribution.
Discussion: Recentering and Rescaling Data

D33. Suppose a U.S. dollar is worth 9.4 Mexican pesos.

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

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

Practice

P32. The mean height of a class of 15 children is 48 inches, the median is 45 inches, the standard deviation is 2.4 inches, and the interquartile range is 3 inches. Find the mean, standard deviation, median, and interquartile range if

a. you convert each height to feet
b. each child grows 2 inches
c. each child grows 4 inches and you convert their heights to feet

P33. Compute means and standard deviations (use the formula for \( s \)) for 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 –900 –1000

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 outliers, it is sensitive to outliers.

Display 2.61 again shows the dot plot for the Nielsen ratings from Display 2.20.
The three highest values—the three shows with the largest numbers of viewers—are outliers.

The printout in Display 2.62 gives summary statistics for all 101 shows.

<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>Ratings</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>
</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>Ratings</td>
<td>2.320</td>
<td>76.260</td>
<td>6.160</td>
<td>12.855</td>
</tr>
</tbody>
</table>

The second printout, in Display 2.63, gives summary statistics when the three outliers are removed from the set of ratings.

<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>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Min</th>
<th>Max</th>
<th>Q1</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Outs</td>
<td>2.320</td>
<td>20.470</td>
<td>6.065</td>
<td>12.698</td>
</tr>
</tbody>
</table>

**Display 2.62** Minitab printout of the summary statistics for all Nielsen ratings.

The second printout, in Display 2.63, gives summary statistics when the three outliers are removed from the set of ratings.

**Display 2.63** Summary statistics for Nielsen ratings without outliers.

**Discussion: The Influence of Outliers**

D34. Are these measures of center affected much by the three outliers? Explain why that is the case.
   a. Mean
   b. Median

D35. Are these measures of spread affected much by the three outliers? Explain why that is the case.
   a. Range
   b. Standard deviation
   c. Interquartile range

**Practice**

P34. The histogram and summary statistics in Display 2.64 and Display 2.65 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 2.65, but substitute your best estimate for the summary statistics now that Hawaii has been excluded.

Display 2.64  Record low temperatures for the states.

<table>
<thead>
<tr>
<th>Summary of</th>
<th>Lowest Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Selector</td>
<td></td>
</tr>
<tr>
<td>Percentile 25</td>
<td></td>
</tr>
<tr>
<td>Count</td>
<td>50</td>
</tr>
<tr>
<td>Mean</td>
<td>-40.3800</td>
</tr>
<tr>
<td>Median</td>
<td>-40</td>
</tr>
<tr>
<td>StdDev</td>
<td>17.6946</td>
</tr>
<tr>
<td>Min</td>
<td>-80</td>
</tr>
<tr>
<td>Max</td>
<td>12</td>
</tr>
<tr>
<td>Range</td>
<td>92</td>
</tr>
<tr>
<td>Lower ith %tile</td>
<td>-51</td>
</tr>
<tr>
<td>Upper ith %tile</td>
<td>-30</td>
</tr>
</tbody>
</table>

Display 2.65  Summary statistics for lowest temperatures by state.

**Summaries from a Frequency Table**

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

\[
\bar{x} = \frac{5 \cdot 6 + 8 \cdot 3}{6 + 3} = \frac{30 + 24}{9} = \frac{54}{9} = 6
\]

You can use formulas to find the mean and standard deviation of a frequency table like the one in Display 2.66.
Formulas for the Mean and Standard Deviation of 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

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

where \( n \) is the sum of the frequencies or \( n = \sum f \).

Example

Suppose you have 5 pennies, 3 nickels, and 2 dimes. Find the mean value per coin and the standard deviation.

Solution

The table in Display 2.66 shows a way to organize the steps for computing the mean using the formula for the mean of a frequency table.

<table>
<thead>
<tr>
<th>Value ( x )</th>
<th>Frequency ( f )</th>
<th>( xf )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Penny</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Nickel</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>Dime</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td><strong>Sum</strong></td>
<td><strong>10</strong></td>
<td><strong>40</strong></td>
</tr>
</tbody>
</table>

\[
\bar{x} = \frac{\sum xf}{n} = \frac{40}{10} = 4
\]

Display 2.66 Steps for computing the mean for a frequency table.

Display 2.67 gives an extended version of the table, designed to organize the steps for computing both the mean and the standard deviation.

<table>
<thead>
<tr>
<th>Value ( x )</th>
<th>Frequency ( f )</th>
<th>( xf )</th>
<th>( x - \bar{x} )</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>
</tr>
<tr>
<td>Nickel</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Dime</td>
<td>10</td>
<td>2</td>
<td>6</td>
<td>36</td>
</tr>
<tr>
<td><strong>Sum</strong></td>
<td><strong>10</strong></td>
<td><strong>40</strong></td>
<td></td>
<td><strong>120</strong></td>
</tr>
</tbody>
</table>

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

Display 2.67 Steps for computing the SD for a frequency table.
Discussion: Summaries from a Frequency Table

D36. Display 2.68 shows the data on family size for two representative sets of 100 families, one set from 1967 and one from 1997.

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 for 1967.

D37. Explain why the formula for the standard deviation in the boxed summary above gives the same answer as the formula on page 45.

Practice

P35. Refer to Display 2.68.

a. Use the formula for the mean and standard deviation of a frequency table to compute the mean number of children per family and the standard deviation for 1997.

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 1967 and compute the IQR. Do the same for 1997.

d. Write a comparison of the distributions for the two years.

P36. Suppose you have 5 pennies, 6 nickels, 4 dimes, and 5 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 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 a frequency table.
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 how many there are.

- 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, add the squared deviations, then 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 shows outliers as well.

An outlier is any value more than \( 1.5 \cdot IQR \) from the nearest quartile. If a summary statistic doesn’t depend much on whether you include or exclude outliers from your data set, then it is said to be resistant.

- The median and quartiles are resistant to outliers.

- The mean and standard deviation, on the other hand, 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 (\( IQR \) and standard deviation) by \( |d| \) and multiplies the measure of center (median and mean) by \( d \).

Exercises

E23. Discuss whether you would use the mean or the median to measure the center of the following sets of data and why you prefer the one you choose.

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
E24. Three histograms and three boxplots appear in Display 2.69. Which boxplot displays the same information as
  a. Histogram A?
  b. Histogram B?
  c. Histogram C?

Display 2.69  Match the histograms with the boxplots.

E25. Make side-by-side boxplots for the speeds of predators and nonpredators. (The stemplot in Display 2.31 shows the values already ordered.) Are the boxplots or the back-to-back stemplot in Display 2.31 better for comparing these speeds? Explain.

E26. The test scores of 40 students in a first-period class were used to construct the first boxplot in Display 2.70, and test scores of 40 students in a second-period class were used for the second. Can the third plot be a boxplot of the scores of the 80 students in the two classes combined? Why or why not?

Display 2.70  Boxplots for two sets of test scores.

E27. 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 seventh value?

E28. No computing should be necessary to answer these questions.
  a. The mean of each of the following 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 the following sets of values have a standard deviation of about 5. Which two are they?
     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

E29. The standard deviation of the first set of values below is about 30. What is the standard deviation of the second set? Explain. No computing should be necessary.

16 23 34 56 78 92 93
20 27 38 60 82 96 97

E30. Consider the set of the heights of all female NCAA athletes and the set of heights of all female NCAA basketball players. Which distribution will have the larger mean? Which will have the larger standard deviation? Explain.

E31. 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 hours, do you need to know the average speed or the median speed in order to get the total distance?
c. Suppose that all trees in a forest are right circular cylinders with a radius of 3 feet. The heights vary, but the mean height is 45 feet, the median is 43 feet, the IQR is 3 feet, and the standard deviation is 3.5 feet. From this information, can you compute the total volume of wood?

E32. Consider the following data set: 15, 8, 25, 32, 14, 8, 25, 2. You may 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?

E33. The histogram in Display 2.71 shows record high temperatures by state.

a. Suppose each of the temperatures is converted from degrees Fahrenheit, F, to degrees Celsius, C, using the formula

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

   Make a histogram of the temperatures in °C.

b. The summary statistics in Display 2.72 are for the temperatures in °F. Make a similar table for the temperatures in °C.

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

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>TrMean</th>
<th>StDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>HighTemp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>100.00</td>
<td>114.10</td>
<td>114.00</td>
<td>113.95</td>
<td>6.69</td>
</tr>
<tr>
<td>Max</td>
<td>134.00</td>
<td>110.00</td>
<td>118.00</td>
<td>118.00</td>
<td></td>
</tr>
</tbody>
</table>

Display 2.72 Summary statistics for record high temperatures for the 50 U.S. states.

E34. Suppose the sum of the squared deviations 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?

E35. This table shows the weights of pennies from Display 2.4 with the weights for each penny taken to be the value at the midpoint of the interval.

<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 of the weights.
c. Does the standard deviation appear to represent a typical deviation from the mean?

E36. For the countries of Europe, many of the average life expectancies are approximately the same, as you can see from the stemplot in Display 2.47. Use the formulas for a frequency table to compute the mean and standard deviation of the life expectancies for the countries of Europe.

E37. 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. Find the medians and quartiles, and use them to write a verbal comparison of the two distributions.

E38. Using only the basic boxplot in Display 2.73, show that there must be outliers in the set of average longevity.

Display 2.73  Boxplot of average longevity.

E39. Display 2.74 shows the boxplot of average longevity, showing outliers. How many outliers are there?

Display 2.74  Modified boxplot of average longevity, showing outliers.

E40. Without computing, what can you say about the standard deviation of this set of values: 4, 4, 4, 4, 4, 4, 4, 4?

E41. Tell how you could use recentering and rescaling to simplify the computation of the mean and standard deviation for this list of numbers:

<table>
<thead>
<tr>
<th>5478.1</th>
<th>5478.3</th>
<th>5478.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>5478.9</td>
<td>5478.4</td>
<td>5478.2</td>
</tr>
</tbody>
</table>

E42. 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 with the formula for the mean of the recentered data.

E43. 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 with that for the standard deviation of the recentered data.

E44. In 1998, 32 of the 50 U.S. states either had no death penalty or executed no one. Of the states that did carry out executions, Texas led the list with 20 executions, followed by Virginia (13); South Carolina (7); Arizona, Oklahoma, and Florida (4 each); and Missouri and North Carolina (3 each). Another 10 states executed 1 person. 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.

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

- It tells you how variability in measuring often behaves (tennis balls).
- It tells you how variability in populations often behaves (weights of pennies, SAT scores).
- It tells you how averages (and some other summary statistics) behave when you repeat a random process (Westvaco case, Activity 1.1).

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 may appear tall and thin while another looks short and fat. However, the x-axis of the tall, thin one can be stretched out so that the two normal distributions look exactly the same.

**Unknown Percentage and Unknown Value Problems**

The basic skills you need in order to utilize the normal distribution are illustrated by solving two related problems: the unknown percentage problem and the unknown value problem. Here’s one of each type.

In a recent year, the distribution of SAT I scores for the incoming class at the University of Washington was roughly normal in shape, with mean 1055 and standard deviation 200.

**Unknown percentage problem** (Display 2.75): What percentage of scores were 920 or below?

*Display 2.75 The unknown percentage problem.*

**Unknown value problem** (Display 2.76): What SAT score separates the lowest 25% of the SAT scores from the rest?
Notice how the two problems are counterparts. To find an unknown percentage, $P$, you must know the corresponding value, $x$. To find an unknown value, you must know the corresponding percentage.

**Discussion: Unknown Percentage and Unknown Value Problems**

D38. Which of the following situations are unknown percentage problems, and which are unknown value problems? For each, draw and label a normal curve, showing the three quantities that are given and the one quantity to find.

a. In the Westvaco simulation of Chapter 1, the averages from 1000 random samples of size 3 were roughly normal, with mean 46.9 and standard deviation 6.1. What is the chance of getting an average of 58 or more?

b. In another set of 1000 random samples, the distribution of averages was also normal, with mean 46.4 and standard deviation 6.2. For this distribution, find the age that cuts off the largest 2.5% of the values.

**Practice**

P37. Which of the following situations are unknown percentage problems, and which are unknown value problems? For each, draw and label a normal curve, showing the three quantities that are given and the one quantity to find.

a. In a recent year, students entering the University of Florida had a mean SAT I score of 1135, with standard deviation 180. The distribution was roughly normal. What percentage of SAT I scores were greater than 1300?

b. In 2000, the mean SAT I math score nationally was 514, with a standard deviation of 113. Find the upper quartile of the distribution.
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 that has mean 0 and standard deviation 1. Solving unknown percentage and unknown value problems depends on this important property.

The normal distribution that has mean 0 and standard deviation 1 is called the standard normal distribution. With this distribution, we call the variable along the horizontal axis a \( z \)-score.

The standard normal distribution is symmetric, with total area under the curve equal to 1, or 100\%. To find the percentage, \( P \), that is the area to the left of the corresponding \( z \)-score, you can use the \( z \)-table or your calculator.

The next two examples show how you use the \( z \)-table, which is Table A in the appendix.

**Example**

Find the percentage, \( P \), of values below \( z = 1.23 \).

**Solution**

Think of 1.23 as 1.2 + .03. In Table A in the appendix, find the row headed 1.2 and the column headed .03. Where this row and column intersect, you find the decimal .8907. So 89.07\% of standard normal scores are below 1.23.
Example

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 the values from the rest.

Solution

Look for .7500 in the body of Table A. No value is exactly equal to .7500. The closest value is .7486, which is close enough. The .7486 sits at the intersection of the row headed 0.6 and the column headed .07, so the corresponding $z$-score is roughly $0.6 + 0.07 = 0.67$.

If you have a graphing calculator, you can find the percentage or value directly. On the TI-83, for example, normalcdf (–99999,1.23) returns a value of .8907, or 89.07%. To find the 75th percentile of a standard normal, use the command invNorm(.75) to get .67449.

Discussion: The Standard Normal Curve

D39. 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?

D40. 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?
2.4 The Normal Distribution

Practice

P38. Find the z-score that has the given percentage of values below it.
   a. 32%   b. 41%   c. 87%   d. 94%

P39. Find the percentage of values below each z-score.
   a. –2.23   b. –1.67   c. –0.40   d. 0.80

P40. What percentage of values in a standard normal distribution fall between
   a. –1.46 and 1.46?
   b. –3 and 3?

P41. For a standard normal distribution, what interval contains
   a. the middle 90% of the z-scores?
   b. the middle 95% of z-scores?

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.

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 of the values by the standard deviation. This gives you a distribution
with standard deviation 1. Now you have a standard normal distribution. You can
also think of the two-step process 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

The distribution of SAT scores for the incoming class at the University of
Washington had mean 1055 and standard deviation 200. What is the z-score for
a University of Washington student who got 912 on the SAT?
Solution

A score of 912 is 143 points below the mean of 1055. This is \(-\frac{143}{200}\) or 0.715 standard deviations below the mean. Alternatively, using the formula,

\[
z = \frac{x - \text{mean}}{SD} = \frac{912 - 1055}{200} = -0.715
\]

so the student’s z-score is \(-0.715\).

Example

What did a student at the University of Washington get on the SAT if his or her score was 1.6 standard deviations above the average?

Solution

The score that is 1.6 standard deviations above average is

\[
x = \text{mean} + z \cdot SD = 1055 + 1.6(200) = 1375
\]

Discussion: Standard Units

D41. Standardizing is a process that is similar to others you have seen already.

a. If you’re driving at 60 mph 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 one corresponds to rescaling?

D42. In the United States, heart disease kills roughly one-and-one-half times as many people as cancer. (Among 100,000 residents, there are 289 deaths per year from heart disease and 200 from cancer.) If you look at these death rates by state, the distributions are roughly normal, provided that you leave out Alaska, which is an outlier. The means and standard deviations are

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart disease</td>
<td>289</td>
<td>54</td>
</tr>
<tr>
<td>Cancer</td>
<td>200</td>
<td>31</td>
</tr>
</tbody>
</table>

Alaska has 90 deaths per 100,000 residents from heart disease, 84 from cancer. Explain which death rate is more extreme compared to other states.

2.4 The Normal Distribution

Practice

P42. Refer to the table in D42. California has 240 deaths from heart disease and 166 deaths from cancer per 100,000 residents. Which rate is more extreme compared to other states, and why?

P43. Refer to the table in D42.

a. Florida has 365 deaths from heart disease and 257 deaths from cancer per 100,000 residents. Which rate is more extreme?

b. Colorado has an unusually low rate of heart disease, 184 deaths per 100,000 residents. Texas has an unusually low rate of cancer, 161 per 100,000 residents. Which is more extreme?

P44. **Standardizing.** Convert each of these values to standard units, \( z \). (Do not use a calculator. These are meant to be done in your head.)

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

P45. **Unstandardizing.** In your head, convert each of these \( z \)-scores back to the scale it came from. That is, find \( x \).

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Solving the Unknown Percentage Problem and Unknown Value Problem

Now you know all you need to solve problems involving any normal distribution.

**For an unknown percentage problem:**

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

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

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,

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

For groups of similar individuals, heights are often 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 inches and standard deviation 2.7 inches. What percentage of these males are more than 74 inches tall?


Solution

Standardize:

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

Look up the percentage: The area to the left of the *z*-score 1.44 is .9251. So the percentage taller than 74 inches is 100 – 92.51 or 7.49%.

Example

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 inches and standard deviation 2.5 inches. What height separates the shortest 75% from the tallest 25%?

Display 2.80  The 75th percentile in height.
Solution

Look up the $z$-score: If the percentage $P = .75$, then from Table A, $z = 0.67$. Unstandardize:

$$x = \text{mean} + z \cdot \text{SD} = 64.8 + 0.67(2.5) \approx 66.475 \text{ inches}$$

Discussion: Solving the Unknown Percentage Problem and the Unknown Value Problem

D43. The heights of 18- to 24-year-old males in the United States are approximately normal with mean 70.1 inches and standard deviation 2.7 inches.

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

b. There are roughly 13,000,000 males between 18 and 24 in the United States. About how many of them are between 67 and 68 inches tall?

c. Find the male height that falls at the 90th percentile.

D44. If the measurements of height are transformed from inches to feet, will that change the shape of the distribution in D43? Describe the distribution of male heights in terms of feet rather than inches.

D45. For 17-year-olds in the United States, blood cholesterol levels in milligrams per deciliter have a normal distribution, approximately, with mean 176 mg/dl and standard deviation 30 mg/dl. The middle 90% of the cholesterol levels are between what two values?

Practice

P46. The heights of 18- to 24-year-old males in the United States are approximately normal with mean 70.1 inches and standard deviation 2.7 inches. The heights of 18- to 24-year-old females have a mean of 64.8 inches and a standard deviation of 2.5 inches.

a. Estimate the percentage of U.S. males between 18 and 24 who are 6 feet tall or taller.

b. How tall does a U.S. woman between 18 and 24 have to be to be at the 35th percentile?

P47. For students entering the University of Florida in a recent year, the distribution of SAT scores was roughly normal, with mean 1100 and standard deviation 180. The middle 95% of the SAT scores were between what two values?
Central Intervals for Normal Distributions

You learned in Section 2.1 that if a distribution is roughly normal, about two-thirds of the values lie within one standard deviation of the mean. (The actual percentage is closer to 68%.) It is helpful to memorize this fact as well as the others in the box that follows.

**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.
Discussion: Central Intervals

D46. Refer to the table in D42.
   a. The middle 90% of the states’ death rates from heart disease fall between what two numbers?
   b. The middle 95% of the death rates from heart disease are between what two numbers?
   c. The middle 90% of the death rates from cancer are between what two numbers?

Practice

P48. Refer to the table in D42. Which of the following rates per 100,000 residents are outside the middle 95% of their distribution? Which of them are outside the middle 99%?
   a. California’s death rate from heart disease of 240.
   b. California’s death rate from cancer of 166.
   c. Alaska’s death rate from heart disease of 90.
   d. Alaska’s death rate from cancer of 84.

Summary 2.4: The Normal Distribution

The standard normal distribution has mean 0, standard deviation 1. The curve is symmetric, with total area of 1. All normal distributions can be converted to the standard normal by converting to standard units:
   • First, recenter by subtracting the mean.
   • Then rescale by dividing by the standard deviation:

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

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

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 1.96 (or about 2) standard deviations of the mean
   • 99.7% (or almost all) of the values lie within 3 standard deviations of the mean
Exercises

E45. On the same set of axes, draw two normal curves with mean 50, one having a standard deviation of 5 and the other having a standard deviation of 10.

E46. 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?

E47. SAT I verbal scores are scaled so that they are approximately normal; the mean is about 505, and the standard deviation is about 111.
   a. Find the probability that a randomly selected student has an SAT I verbal score
      i. between 400 and 600
      ii. over 700
      iii. below 450
   b. What SAT I verbal scores fall in the middle 95% of the distribution?

E48. SAT I math scores are scaled so that they are approximately normal and the mean is about 511 and the standard deviation is 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?

E49. Height limitations for flight attendants. To work as a flight attendant for United Airlines, you must be between 5’2” and 6’ tall. [Source: www.ual.com/airline.] The mean height of 18- to 24-year-old males in the United States is about 70.1 inches, with a standard deviation of 2.7 inches. The mean height of 18- to 24-year-old females is about 64.8 inches, with a standard deviation of 2.5 inches. Both distributions are approximately normal. What percentage of men this age meet this height limitation? What percentage of women this age meet this height limitation?

E50. Where is the next generation of male professional basketball players coming from?
   a. Use the normal distribution to approximate the percentage of men in the United States between the ages of 18 and 24 who are as tall or taller than each basketball player below. Then, using the fact that there are about 13,000,000 men between 18 and 24 in the United States, estimate how many are as tall or taller than each player.
      i. Karl Malone, 6’9”
      ii. Michael Jordan, 6’6”
      iii. Shaquille O’Neal, 7’1”
   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 parts i–iii are too small, too big, or just right?

E51. 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.)

E52. The British monarchy. Over the 1200 years of the British monarchy, the average reign of kings and queens 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 the distribution can possibly be normal-shaped.
E53. **NCAA scores.** The histogram in Display 2.81 was constructed from the total of the scores of both teams in all NCAA basketball play-off games between 1939 and 1995.

a. Approximate the mean of this distribution.

b. Approximate the standard deviation of this distribution.

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

d. Suppose you choose a game at random from next year’s 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?

Display 2.81 Total points scored in NCAA play-off games.


E54. **Puzzle problems.** In problems that involve computations with the normal distribution, there are four quantities: mean, standard deviation, value \( x \), and percentage \( P \) below value \( x \). Any three are enough to determine the fourth. Think of each row in this table as little puzzles and find the missing value in each case. They aren’t the sort of thing you are likely to run into in practice, but solving them can help you become more skilled using the normal distribution.

<table>
<thead>
<tr>
<th>Mean</th>
<th>Standard Deviation</th>
<th>( x )</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>–?–</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>–?–</td>
<td>.18</td>
</tr>
<tr>
<td>–?–</td>
<td>3</td>
<td>6</td>
<td>.09</td>
</tr>
<tr>
<td>10</td>
<td>–?–</td>
<td>12</td>
<td>.60</td>
</tr>
</tbody>
</table>

E55. **More puzzle problems.** In each row below, assume the distribution is normal. Since 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>Standard Deviation</th>
<th>( Q_1 )</th>
<th>( Q_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>5</td>
<td>–?–</td>
<td>–?–</td>
</tr>
<tr>
<td>–?–</td>
<td>–?–</td>
<td>120</td>
<td>180</td>
</tr>
<tr>
<td>–?–</td>
<td>10</td>
<td>100</td>
<td>–?–</td>
</tr>
<tr>
<td>10</td>
<td>–?–</td>
<td>–?–</td>
<td>11</td>
</tr>
</tbody>
</table>

E56. For the comparisons below, you will be using either the SAT I verbal scores in Display 2.55 or assuming that the scores have a normal distribution with mean 505 and standard deviation 111.

a. Estimate the percentile for an SAT I verbal score of 425 using the plot. Then find the percentile for a score of 425 using a \( z \)-score. Are the two values close?

b. Estimate the SAT I verbal score that falls at the 40th percentile using the table in Display 2.55. Then find the 40th percentile using a \( z \)-score. Are they close?

c. Estimate the median from the cumulative relative frequency plot. Is this 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.
Chapter Summary: Exploring Distributions

Distributions come in various shapes, and the appropriate summaries (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 normal-shaped 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 the $z$-table to find the percentage of values in any interval.

The mean and standard deviation are not resistant—their values are sensitive to outliers. So for skewed distributions, you ordinarily use the median (halfway point) and quartiles (medians of the lower and upper halves of the data) as summaries.

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

Review Exercises

E57. The map in Display 2.82, from the U.S. National Weather Service, gives the number of tornadoes by state, including the District of Columbia.

a. Construct a stemplot of the number of tornadoes.

b. Find 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.

Display 2.82 The number of tornadoes per state in 1995.

E58. The summary of SAT I scores at the University of Michigan indicates that “the middle 50% of the scores were between 1170 and 1340, with half the scores above 1210 and half below.” What SAT I scores would be considered outliers for that university?

E59. The cumulative relative frequency plot in Display 2.83 shows the amount of change carried by a group of 200 students.

```
Display 2.83  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 data skewed right, skewed left, or symmetric?
d. Does the data set look like it should be modeled by a normal distribution? Explain your reasoning.

E60. Display 2.84 shows two sets of graphs. The first set is repeated from P5, and 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.

```
Display 2.84  Four distributions with different shapes and their cumulative relative frequency plots.
```

be in a lower percentile on Test II than on Test I. These computations 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, 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?

E62. The average income in dollars of people in each of the 50 states was computed for 1980 and for 1994. Summary statistics for these two distributions are given in Display 2.85.

<table>
<thead>
<tr>
<th></th>
<th>1980</th>
<th>1994</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>9,594</td>
<td>21,078</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>1,406</td>
<td>3,130</td>
</tr>
<tr>
<td>Minimum</td>
<td>6,926</td>
<td>15,794</td>
</tr>
<tr>
<td>Lower quartile</td>
<td>8,348</td>
<td>18,810</td>
</tr>
<tr>
<td>Median</td>
<td>9,723</td>
<td>20,582</td>
</tr>
<tr>
<td>Upper quartile</td>
<td>10,612</td>
<td>22,542</td>
</tr>
<tr>
<td>Maximum</td>
<td>13,835</td>
<td>30,721</td>
</tr>
</tbody>
</table>


a. Explain the meaning of the $6,926 for the minimum in 1980.

b. Are any states outliers for either year?

c. In 1994 the average personal income in Alabama was $17,926, and in 1980 it was $7,704. Did the income in Alabama change much in relation to the other states? Explain your reasoning.

E63. The *World Almanac* records high and low temperatures by state from 1890 through 1999. Stem-and-leaf plots of the years that each state had its lowest temperature and the years that each state had its highest temperature appear 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.

<table>
<thead>
<tr>
<th></th>
<th>1980</th>
<th>1994</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>9,594</td>
<td>21,078</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>1,406</td>
<td>3,130</td>
</tr>
<tr>
<td>Minimum</td>
<td>6,926</td>
<td>15,794</td>
</tr>
<tr>
<td>Lower quartile</td>
<td>8,348</td>
<td>18,810</td>
</tr>
<tr>
<td>Median</td>
<td>9,723</td>
<td>20,582</td>
</tr>
<tr>
<td>Upper quartile</td>
<td>10,612</td>
<td>22,542</td>
</tr>
<tr>
<td>Maximum</td>
<td>13,835</td>
<td>30,721</td>
</tr>
</tbody>
</table>

*Display 2.86* Stem-and-leaf plots of record low and high temperatures by state.


E64. Display 2.87 shows some results of the Third International Mathematics and Science study for various countries. Each case is a school.

E65. The World Almanac records high and low temperatures by state from 1890 through 1999. Stem-and-leaf plots of the years that each state had its lowest temperature and the years that each state had its highest temperature appear 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.

Display 2.87 Boxplots of mathematics instruction time by country for upper grade 9-year-olds.

a. Find the boxplot for the United States. What, exactly, are the individual values that are plotted?
b. Why are there only lines and not boxes for Norway and Singapore?
c. Describe how the distribution for the United States compares to the distributions for the other countries.

E65. The side-by-side boxplots in Display 2.88 give the percentage of 4th-grade-aged 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 for any region skewed left? skewed right? symmetrical?

d. Did this happen? Which type of plot gives a better impression of the distributions?

E66. 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 2.90. 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>

E67. The average number of pedestrian deaths annually for 41 metropolitan areas is given in Display 2.91.

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

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

E68. A game invented recently by three college students involves giving the name of an actor or actress and then trying to connect that actor or actress 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* (1995) with Marc Macaulay, who was in *Wild Things* (1998) with Kevin Bacon. Display 2.92 gives the number of links required to connect each of the 263,484 actors and actresses in the Internet Movie Database to Kevin Bacon.

<table>
<thead>
<tr>
<th>Metro Area</th>
<th>Average Annual Fatalities</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>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metro Area</th>
<th>Average Annual Fatalities</th>
</tr>
</thead>
<tbody>
<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 2.91 Average annual pedestrian deaths.*

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 2.93.

<table>
<thead>
<tr>
<th>Connery Number</th>
<th>Number of Actors/Actresses</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1,801</td>
</tr>
<tr>
<td>2</td>
<td>113,571</td>
</tr>
<tr>
<td>3</td>
<td>135,546</td>
</tr>
<tr>
<td>4</td>
<td>11,591</td>
</tr>
<tr>
<td>5</td>
<td>824</td>
</tr>
<tr>
<td>6</td>
<td>130</td>
</tr>
<tr>
<td>7</td>
<td>20</td>
</tr>
</tbody>
</table>

Display 2.93 Connery numbers.

E70. How good are batters in the National League? Display 2.95 shows the distribution of batting averages for National Leaguers who batted 100 times or more in 1998.

a. Approximate the mean and standard deviation of the batting averages from the histogram. Find the central interval that contains 95% of the batting averages. Would it be unusual to see a player hitting under .200 (or 200) in the National League?

b. Compare the distributions of batting averages for the two leagues. (See E69 for the American League.) What are the main differences between the two distributions?

E69. How good are the batters in the American League of Major League Baseball? Display 2.94 shows the distribution of batting averages for all American League players who batted 100 times or more in the 1998 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.)

a. Does it look like batting averages are approximately normally distributed?

b. Approximate the mean and standard deviation of the batting averages from the histogram.

c. Give the interval that contains the middle 95% of the batting averages for the American League.

Display 2.94 American League batting averages, 1998.

Chapter 2: Exploring Distributions

c. A batter hitting .200 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?


E71. The statistics below summarize the set of Nielsen ratings from a week without any special programming. You can find a dot plot of the ratings in Display 2.22, E10, at the end of Section 2.1.

<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>Total</td>
<td>113</td>
<td>6.867</td>
<td>6.900</td>
<td>6.596</td>
<td>3.490</td>
<td>0.328</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>Total</td>
<td>1.400</td>
<td>20.700</td>
<td>4.550</td>
<td>8.250</td>
</tr>
</tbody>
</table>

a. Use the dot plot and summary statistics for all 113 shows to make boxplots of the ratings, showing any outliers.

b. Use the summary statistics below to make a set of side-by-side boxplots for the six networks. You will not be able to show any outliers.

<table>
<thead>
<tr>
<th>Network</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>TrMean</th>
<th>StDev</th>
<th>SEMean</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC</td>
<td>22</td>
<td>7.755</td>
<td>7.750</td>
<td>7.715</td>
<td>2.439</td>
<td>0.520</td>
</tr>
<tr>
<td>CBS</td>
<td>25</td>
<td>7.856</td>
<td>7.500</td>
<td>7.765</td>
<td>1.728</td>
<td>0.346</td>
</tr>
<tr>
<td>FOX</td>
<td>16</td>
<td>6.156</td>
<td>6.050</td>
<td>6.064</td>
<td>1.508</td>
<td>0.377</td>
</tr>
<tr>
<td>NBC</td>
<td>28</td>
<td>9.157</td>
<td>7.100</td>
<td>8.992</td>
<td>4.308</td>
<td>0.814</td>
</tr>
<tr>
<td>UPN</td>
<td>8</td>
<td>2.188</td>
<td>2.150</td>
<td>2.188</td>
<td>0.449</td>
<td>0.159</td>
</tr>
<tr>
<td>WB</td>
<td>14</td>
<td>2.614</td>
<td>2.700</td>
<td>2.617</td>
<td>0.684</td>
<td>0.183</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Network</th>
<th>Min</th>
<th>Max</th>
<th>Q1</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC</td>
<td>2.900</td>
<td>13.400</td>
<td>6.025</td>
<td>9.125</td>
</tr>
<tr>
<td>FOX</td>
<td>4.000</td>
<td>9.600</td>
<td>4.850</td>
<td>6.900</td>
</tr>
<tr>
<td>NBC</td>
<td>4.500</td>
<td>20.700</td>
<td>6.800</td>
<td>11.100</td>
</tr>
<tr>
<td>UPN</td>
<td>1.700</td>
<td>3.100</td>
<td>1.825</td>
<td>2.375</td>
</tr>
<tr>
<td>WB</td>
<td>1.400</td>
<td>3.800</td>
<td>2.075</td>
<td>3.125</td>
</tr>
</tbody>
</table>

E72. A distribution is symmetric with approximately equal mean and median. Is it necessarily the case that about two-thirds of the values are within one standard deviation of the mean? If yes, explain why. If not, give an example.

E73. Construct a set of data where all values are larger than 0, but one standard deviation below the mean is less than 0.

E74. The boxplots in Display 2.96 show the life expectancies for the countries of Africa, Europe, and the Middle East. The table below the plots 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).

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean</th>
<th>Median</th>
<th>StdDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>73.61</td>
<td>73.85</td>
<td>3.05</td>
</tr>
<tr>
<td>B</td>
<td>53.59</td>
<td>51.71</td>
<td>8.37</td>
</tr>
<tr>
<td>C</td>
<td>69.86</td>
<td>70.70</td>
<td>5.22</td>
</tr>
</tbody>
</table>

Display 2.96 Life expectancies for the countries of Africa, Europe, and the Middle East.

E75. Another measure of center that is sometimes used is the midrange. To find the midrange, compute the mean of the largest value and the smallest value.

The statistics in the computer output given here summarize the Nielsen data for the week of the last new Seinfeld episode.
a. Using these summary statistics alone, compute the midrange both with and without the Seinfeld episode. (You will need to refer to Display 2.20 on page 17 to get one number.) Is the midrange sensitive to outliers, or is it resistant to outliers? Explain.

b. Compute the mean of the ratings without the Seinfeld episode using only the summary statistics above.

E76. In computer output like that of E75, the TrMean is the **trimmed mean**. It typically is computed by removing the largest 5% of the values and the smallest 5% of the values from the data set and then computing the mean of the remaining middle 90% of the 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.

b. Is the trimmed mean resistant to outliers?