



Describing Distributions with Numbers

In Chapter 1 we began to learn about data analysis using graphs to provide a visual tool for organizing and identifying patterns in data. Pie charts and bar graphs can summarize the information in a categorical variable by giving us the percentage of the distribution in the various categories. Histograms and stemplots are graphical tools for summarizing the information provided by a quantitative variable. The overall pattern in a histogram or stemplot illustrates some of the important features of the distribution of a variable. The center of the histogram tells us about the value of a “typical” observation on this variable, whereas the variability gives us a sense of how close most of the observations are to this value. Other interesting features are the presence of outliers and the general shape of the plot. For data collected over time, time plots can show patterns such as seasonal variation and trends in the variable. In the next chapter, we will see how the information about the distribution of a variable can also be described using numerical summaries.

In this chapter, we continue our study of exploratory data analysis. A graph is an important visual tool for organizing and identifying patterns in data. It gives a fairly complete description of a distribution, although for many problems, the important information in the data can be described by using a few numbers. A graph is a numerical summary that can be useful for describing a single distribution as well as for comparing the distributions from several groups of observations.

Brasil2/Getty Images

CHAPTER

2

When you complete this chapter, you will be able to:

- 2.1 Find the mean of a set of observations and interpret it as the center of the set.
- 2.2 Find the median of a set of observations and interpret it as the center of the set.
- 2.3 Compare the mean and median values of a data set and distinguish their meanings.
- 2.4 Calculate and use quartiles to describe the variation of a data set.
- 2.5 Use the five-number summary (the minimum, the maximum, the quartiles, and the median) and boxplot to characterize a distribution.
- 2.6 Use the $1.5 \times IQR$ rule to identify outliers in a data set.
- 2.7 Calculate and use the standard deviation to describe the variation of a data set.
- 2.8 Discriminate between the five-number summary and the mean and standard deviation for describing the distribution of data, depending on features of the data set.
- 2.9 Interpret presentations of descriptive statistics output by graphing calculators and computer programs.
- 2.10 Apply four-step state, plan, solve, and conclude process to examine data based on context.

We saw in Chapter 1 (page 13) that the American Community Survey asks, among much else, about workers' travel times to work. Here are the travel times in minutes for 15 workers in North Carolina, chosen at random by the U.S. Census Bureau:¹

20 35 8 70 5 15 25 30 40 35 10 12 40 15 20

We aren't surprised that most people estimate their travel time in multiples of five minutes. Here is a stemplot of these data:

```

0 | 5 8
1 | 0 2 5 5
2 | 0 0 5
3 | 0 5 5
4 | 0 0
5 |
6 |
7 | 0

```

The distribution is single peaked and right-skewed. The longest travel time (70 minutes) may be an outlier. Our goal in this chapter is to describe with numbers the center and variability of this and other distributions.

2.1 Measuring Center: The Mean

The most common measure of center is the ordinary arithmetic average, or *mean*.

The Mean \bar{x}

To find the **mean** of a set of observations, add their values and divide by the number of observations. If the n observations are x_1, x_2, \dots, x_n , their mean is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

or, in more compact notation,

$$\bar{x} = \frac{1}{n} \sum x_i$$

The Σ (capital Greek sigma) in the formula for the mean is short for “add them all up.” The subscripts on the observations x_i are just a way of keeping the n observations distinct. They do not necessarily indicate order or any other special facts about the data. The bar over the x indicates the mean of all the x -values. Pronounce the mean \bar{x} as “x-bar.” This notation is very common. When writers who are discussing data use \bar{x} or \bar{y} , they are talking about a mean.

EXAMPLE 2.1 Travel Times to Work

The mean travel time of our 15 North Carolina workers is

$$\begin{aligned}
 \bar{x} &= \frac{x_1 + x_2 + \cdots + x_n}{n} \\
 &= \frac{20 + 35 + \cdots + 20}{15} \\
 &= \frac{380}{15} = 25.3 \text{ minutes}
 \end{aligned}$$



In practice, you can enter the data into your calculator and ask for the mean. You don't have to actually add and divide. But you should know that this is what the calculator is doing.

Notice that only 6 of the 15 travel times are larger than the mean. If we leave out the longest single travel time, 70 minutes, the mean for the remaining 14 people is 22.1 minutes. That one 70-minute observation raises the mean by 3.2 minutes.

Example 2.1 illustrates an important fact about the mean as a measure of center: it is sensitive to the influence of a few extreme observations. These may be outliers, but a skewed distribution that has no outliers will also pull the mean toward its long tail. Because the mean cannot resist the influence of extreme observations, we say that it is not a *resistant measure* of center.

Resistant Measure

A **resistant measure** is a statistical measure that is relatively unaffected by large changes in numerical values of a small proportion of the observations in the distribution that the measure describes.

APPLY YOUR KNOWLEDGE

2.1 E. Coli in Swimming Areas. To investigate water quality, the *Columbus Dispatch* took water specimens at 16 Ohio State Park swimming areas in central Ohio. Those specimens were taken to laboratories and tested for *E. coli*, which are bacteria that can cause serious gastrointestinal problems. For reference, if a 100-milliliter specimen (about 3.3 ounces) of water contains more than 130 *E. coli* bacteria, it is considered unsafe. Here are the *E. coli* levels per 100 milliliters found by the laboratories:²

291.0	10.9	47.0	86.0	44.0	18.9	1.0	50.0
190.4	45.7	28.5	18.9	16.0	34.0	8.6	9.6

Find the mean *E. coli* level. How many of the lakes have *E. coli* levels greater than the mean? What feature of the data explains the fact that the mean is greater than most of the observations?

2.2 Health Care Spending. Table 1.3 (page 31) gives the 2015 health care expenditure per capita in 35 countries with the highest gross domestic products in 2015. The United States, at 9536 international dollars per person, is a high outlier. Find the mean health care spending in these nations with and without the United States. How much does the one outlier increase the mean?

STATISTICS IN YOUR WORLD

Don't Hide the Outliers

Data from an airliner's control surfaces, such as the vertical tail rudder, go to cockpit instruments and then to the "black box" flight data recorder. To avoid confusing the pilots, short erratic movements in the data are "smoothed" so that the instruments show overall patterns. When a crash killed 260 people, investigators suspected a catastrophic movement of the tail rudder. But the black box contained only the smoothed data. Sometimes, outliers are more important than the overall pattern.


2.2 Measuring Center: The Median

In Chapter 1, we used the midpoint of a distribution as an informal measure of center and gave a method for its computation. The *median* is the formal version of the midpoint, and we now provide a more detailed rule for its calculation.

The Median M

The **median M** is the midpoint of a distribution, the number such that half the observations are smaller and the other half are larger. To find the median of a distribution:

1. Arrange all observations in order of size, from smallest to largest.
2. If the number of observations n is odd, the median M is the center observation in the ordered list. If the number of observations n is even, the median M is midway between the two center observations in the ordered list.
3. You can always locate the median in the ordered list of observations by counting up $(n + 1)/2$ observations from the start of the list.

 Note that the formula $(n + 1)/2$ does not give the median; it just gives the location of the median in the ordered list. Medians require little arithmetic, so they are easy to find by hand for small sets of data. Arranging even a moderate number of observations in order is very tedious, however, so finding the median by hand for larger sets of data is unpleasant. Even simple calculators have an \bar{x} button, but you will need to use software or a graphing calculator to automate finding the median.

EXAMPLE 2.2 Finding the Median: Odd n



What is the median travel time for our 15 North Carolina workers? Here are the data arranged in order:

5 8 10 12 15 15 20 20 25 30 35 35 40 40 70

The count of observations, $n = 15$, is odd. The bold 20 is the center observation in the ordered list, with seven observations to its left and seven to its right. This is the median, $M = 20$ minutes.

Because $n = 15$, our rule for the location of the median gives

$$\text{location of } M = \frac{n + 1}{2} = \frac{16}{2} = 8$$

That is, the median is the eighth observation in the ordered list. It is more reliable and faster to use this rule than to locate the center by eye.

EXAMPLE 2.3 Finding the Median: Even n



Travel times to work in New York State are (on the average) longer than in North Carolina. Here are the travel times in minutes of 20 randomly chosen New York workers:

10 15 55 20 65 50 12 20 10 10 35 50 30 45 15 10 75 40 35 60

A stemplot not only displays the distribution but also makes finding the median easy because it arranges the observations in order:

0	
1	0000255
2	00
3	055
4	05
5	005
6	05
7	5

The distribution is single peaked and right-skewed, with several travel times of an hour or more. There is no center observation, but there is a center pair. These are the bold 30 and 35 in the stemplot, which have nine observations before them in the ordered list and nine after them. The median is midway between these two observations:

$$M = \frac{30 + 35}{2} = 32.5 \text{ minutes}$$

With $n = 20$, the rule for locating the median in the list gives

$$\text{location of } M = \frac{n+1}{2} = \frac{21}{2} = 10.5$$

The location 10.5 means “halfway between the 10th and 11th observations in the ordered list.” That agrees with what we found by eye.

2.3 Comparing the Mean and the Median

Examples 2.1 and 2.2 illustrate an important difference between the mean and the median. The median travel time (the midpoint of the distribution) is 20 minutes. The mean travel time is higher, 25.3 minutes. The mean is pulled toward the right tail of this right-skewed distribution. The median, unlike the mean, is *resistant*. If the longest travel time were 700 minutes rather than 70 minutes, the mean would increase to 67.3 minutes, but the median would not change at all. The outlier just counts as one observation above the center, no matter how far above the center it lies. The mean uses the actual value of each observation and so will chase a single large observation upward. Using the *Mean and Median* applet is an excellent way to compare the resistance of M and \bar{x} .



Comparing the Mean and the Median

The mean and median of a roughly symmetric distribution are close together. If the distribution is exactly symmetric, the mean and median are exactly the same. In a skewed distribution, the mean is usually farther out in the long tail than the median.³

Many economic variables have distributions that are skewed to the right. For example, the median endowment of colleges and universities in the United States and Canada in 2018 was about \$142 million—but the mean endowment was over \$770 million. Most institutions have modest endowments, but a few are very wealthy. Harvard’s endowment was more than \$38 billion.⁴ The few wealthy institutions pull the mean up but do not affect the median. Reports about incomes and other strongly skewed distributions usually give the median (“midpoint”) rather than the mean (“arithmetic average”). However, a county that is about to impose a tax of 1% on the incomes of its residents cares about the mean income, not the median. The tax revenue will be 1% of total income, and because the total income is the mean income times the number of residents, the tax revenue can be computed easily from the mean. The mean and median measure center in different ways, and both are useful. *Don’t confuse the “average” value of a variable (the mean) with its “middle”*




value, which we might describe by the median, or with its “typical” value, which we might describe by the mode.


The **mode** of a set of values is the most frequently occurring value. The mode can be calculated for both numerical and categorical data. In Example 1.3, the mode of the audio brands is Spotify because the highest percentage of 12- to 34-years-olds have listened to this brand. For the 20 New York travel times, the mode is 10 minutes because this time occurred most often in the 20 travel times. There can be multiple modes; for example, in the 15 North Carolina travel times, 15, 20, 35, and 40 minutes are all modes because all these numbers are tied for occurring most often.

APPLY YOUR KNOWLEDGE

- 2.3

New York Travel Times. Find the mean of the travel times to work for the 20 New York workers in Example 2.3. Compare the mean and median for these data. What general fact does your comparison illustrate?  NYTRAVEL
- 2.4

New House Prices. The mean and median sales prices of new homes sold in the United States in July 2019 were \$312,800 and \$388,000.⁵ Which of these numbers is the mean, and which is the median? Explain how you know.
- 2.5


Carbon Dioxide Emissions. Burning fuels in power plants and motor vehicles emits carbon dioxide (CO₂), which contributes to global warming. The CO₂ emissions (metric tons per capita) for countries varies from 0.04 in Burundi to 43.86 in Qatar. Although the data set includes 203 countries, the CO₂ emissions of 14 countries are not available on the World Bank database. The data set is too large to print here, but here are the data for the first 5 countries:⁶  CO2EMISS



Country	CO ₂ Emissions (metric tons per capita)
Aruba	8.41
Afghanistan	0.29
Angola	1.29
Albania	1.98
Andorra	5.83

Find the mean and the median for the full data set (included among the data sets available for this chapter). Make a histogram of the data. What features of the distribution explain why the mean is larger than the median?

2.4 Measuring Variability: The Quartiles

The mean and median provide two different measures of the center of a distribution. But a measure of center alone can be misleading. The U.S. Census Bureau reports that in 2017 the median income of American households was \$61,372. Half of all households had incomes below \$61,372, and half had higher incomes. The mean was much higher, \$86,220, because the distribution of incomes is skewed to the right. But the median and mean don't tell the whole story. The bottom 20% of households had incomes less than \$24,638, and households in the top 5% took in more than \$237,034.⁷ We are interested in the *variability* of incomes as well as their center.  The simplest useful numerical description of a distribution requires both a measure of center and a measure of variability.

One way to measure variability is to give the smallest and largest observations. For example, the travel times of our 15 North Carolina workers range from 5 minutes to 70 minutes. These single observations show the full variability of the data, but they may be outliers. We can improve our description of variability by also looking at the variability of the middle half of the data. The *quartiles* mark out the middle half. Count up the ordered list of observations, starting from the smallest. The *first quartile* lies one-quarter of the way up the list. The *third quartile* lies three-quarters of the way up the list. In other words, the first quartile is larger than 25% of the observations, and the third quartile is larger than 75% of the observations. The second quartile is the median, which is larger than 50% of the observations. That is the idea of quartiles. We need a rule to make the idea exact. The rule for calculating the quartiles uses the rule for the median.

The Quartiles Q_1 and Q_3

To calculate the **quartiles**:

1. Arrange the observations in increasing order and locate the median, M , in the ordered list of observations.
2. The **first quartile**, Q_1 , is the median of the observations whose position in the ordered list is to the left of the location of the overall median.
3. The **third quartile**, Q_3 , is the median of the observations whose position in the ordered list is to the right of the location of the overall median.

The following examples show how the rules for the quartiles work for both odd and even numbers of observations.

EXAMPLE 2.4 Finding the Quartiles: Odd n

Our North Carolina sample of 15 workers' travel times, arranged in increasing order, is

5 8 10 12 15 15 20 20 25 30 35 35 40 40 70

There is an odd number of observations, so the median is the middle one, the bold 20 in the list. The first quartile is the median of the seven observations to the left of the median. This is the fourth of these seven observations, so $Q_1 = 12$ minutes. If you want, you can use the rule for the location of the median with $n = 7$:

$$\text{location of } Q_1 = \frac{n+1}{2} = \frac{7+1}{2} = 4$$

The third quartile is the median of the seven observations to the right of the median, $Q_3 = 35$ minutes. *When there is an odd number of observations, leave out the overall median when you locate the quartiles in the ordered list.*

The quartiles are *resistant* because they are not affected by a few extreme observations. For example, Q_3 would still be 35 if the outlier were 700 rather than 70.



EXAMPLE 2.5 Finding the Quartiles: Even n


Here are the travel times to work of the 20 New Yorker workers from Example 2.3, arranged in increasing order:

10 10 10 10 12 15 15 20 20 30 | 35 35 40 45 50 50 55 60 65 75



There is an even number of observations, so the median lies midway between the middle pair, the 10th and 11th in the list. Its value is $M = 32.5$ minutes. We have marked the location of the median by |. The first quartile is the median of the first 10 observations because these are the observations to the left of the location of the median. Check that $Q_1 = 13.5$ minutes and $Q_3 = 50$ minutes. *When the number of observations is even, include all the observations when you locate the quartiles.*

Be careful when, as in these examples, several observations take the same numerical value. Write down all the observations, arrange them in order, and apply the rules just as if they all had distinct values.

 There are several rules for finding the quartiles. Some calculators and software use rules that give results that differ from ours for some sets of data (see Example 2.8). Our rule is the simplest for hand calculation, with the results from the various rules generally being close to each other.

2.5 The Five-Number Summary and Boxplots

The smallest and largest observations tell us little about a distribution as a whole, but they give information about the tails of the distribution that is missing if we know only the median and the quartiles. To get a quick summary of both center and variability, combine all five numbers.

The Five-Number Summary

The **five-number summary** of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is

Minimum Q_1 M Q_3 Maximum

These five numbers offer a reasonably complete description of center and variability. The five-number summaries of travel times to work from Examples 2.4 and 2.5 are

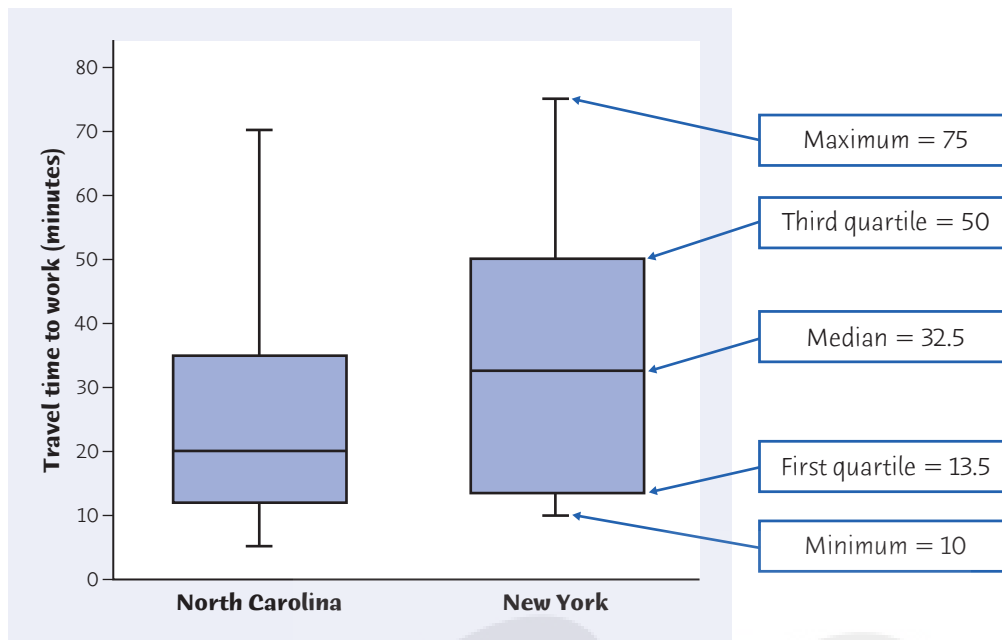
North Carolina	5	12	20	35	70
New York	10	13.5	32.5	50	75

The five-number summary of a distribution leads to a new graph, the *boxplot*. Figure 2.1 shows boxplots comparing travel times to work in North Carolina and New York.

Boxplot

A **boxplot** is a graph of the five-number summary:

- A central box spans the quartiles Q_1 and Q_3 .
- A line in the box marks the median M .
- Lines extend from the box out to the smallest and largest observations.

**FIGURE 2.1**

Boxplots comparing the travel times to work of samples of workers in North Carolina and New York.

Because boxplots show less detail than histograms or stemplots, they are best used for side-by-side comparison of more than one distribution, as in Figure 2.1. Be sure to include a numerical scale in the graph. When you look at a boxplot, first locate the median, which marks the center of the distribution. Then look at the variability. The span of the central box shows the variability of the middle half of the data, and the extremes (the smallest and largest observations) show the variability of the entire data set. We see from Figure 2.1 that travel times to work are in general a bit longer in New York than in North Carolina. The median, both quartiles, the minimum, and the maximum are all larger in New York. New York travel times are also more variable, as shown by the span of the box. Note that the boxes with arrows in Figure 2.1 that indicate the location of the five-number summary are *not* part of the boxplot but are included purely for illustration.

Finally, the North Carolina data are more strongly right-skewed. In a symmetric distribution, the first and third quartiles are equally distant from the median. In most distributions that are skewed to the right, on the other hand, the third quartile is farther above the median than the first quartile is below it. The extremes behave the same way, but remember that they are just single observations and may say little about the distribution as a whole.

APPLY YOUR KNOWLEDGE

2.6 Shared Pain and Bonding. Although painful experiences are involved in social rituals in many parts of the world, little is known about the social effects of pain. Will sharing painful experiences in a small group lead to greater bonding of group members than sharing a similar non-painful experience? Fifty-four university students in South Wales were divided at random into a pain group containing 27 students, with the remaining students in the no-pain group. Pain was induced by two tasks. In the first task, students submerged

their hands in freezing water for as long as possible, moving metal balls at the bottom of the vessel into a submerged container; in the second task, students performed a standing wall squat with back straight and knees at 90 degrees for as long as possible. The no-pain group completed the first task using room temperature water for 90 seconds and the second task by balancing on one foot for 60 seconds, changing feet if necessary. In both the pain and non-pain settings, the students completed the tasks in small groups, which typically consisted of four students and contained similar levels of group interaction. Afterward, each student completed a questionnaire to create a bonding score based on answers to questions such as “I feel the participants in this study have a lot in common,” or “I feel I can trust the other participants.” Here are the bonding scores for the two groups:⁸

No-pain group:	3.43	4.86	1.71	1.71	3.86	3.14	4.14	3.14	4.43	3.71
	3.00	3.14	4.14	4.29	2.43	2.71	4.43	3.43	1.29	1.29
	3.00	3.00	2.86	2.14	4.71	1.00	3.71			
Pain group:	4.71	4.86	4.14	1.29	2.29	4.43	3.57	4.43	3.57	3.43
	4.14	3.86	4.57	4.57	4.29	1.43	4.29	3.57	3.57	3.43
	2.29	4.00	4.43	4.71	4.71	2.14	3.57			

- (a) Find the five-number summaries for the pain and no-pain groups.
- (b) Construct a comparative boxplot for the two groups following the model of Figure 2.1. It doesn’t matter if your boxplots are horizontal or vertical, but they should be drawn on the same set of axes.
- (c) Which group tends to have higher bonding scores? Is the variability in the two groups similar, or does one of the groups tend to have less variable bonding scores? Does either group contain one or more clear outliers?

2.7 Fuel Economy for Midsize Cars. The Department of Energy provides fuel economy ratings for all cars and light trucks sold in the United States. Here are the estimated miles per gallon for combined city and highway driving for the 189 cars classified as midsize in 2019, arranged in increasing order:⁹

12 14 16 16 16 17 17 17 18 18 18 18 19 19 19 19 19 19
20 20 20 20 20 20 20 20 21 21 21 21 21 21 22 22 22 22
22 23 23 23 23 23 23 23 23 23 23 23 23 23 24 24 24 24
24 24 24 24 24 24 24 24 24 24 24 24 24 25 25 25 25 25
25 25 25 25 25 25 25 25 25 25 25 25 26 26 26 26 26 26
26 26 26 26 26 26 26 26 26 26 26 26 26 26 27 27 27 27
27 27 27 27 27 27 27 27 27 27 27 27 27 28 28 28 28 28
29 29 29 29 29 29 29 29 29 29 29 30 30 30 30 30 31 31
31 31 32 32 32 32 32 32 32 32 32 32 32 33 33 33 33 33
33 34 34 34 34 35 35 35 35 36 41 41 41 41 42 42 43 44
44 46 46 48 50 52 52 52 56

- (a) Give the five-number summary of this distribution.
- (b) Draw a boxplot of these data. What is the shape of the distribution shown by the boxplot? Which features of the boxplot led you to this conclusion? Are any observations unusually small or large?

2.6 Spotting Suspected Outliers and Modified Boxplots*

Look again at the stemplot of travel times to work in North Carolina in Example 2.3. The five-number summary for this distribution is

5 12 20 35 70

How shall we describe the variability of this distribution? The smallest and largest observations are extremes that don't describe the variability of the majority of the data. The distance between the quartiles (the range of the center half of the data) is a more resistant measure of variability. This distance is called the *interquartile range*.

The Interquartile Range (IQR)

The **interquartile range (IQR)** is the distance between the first and third quartiles,

$$IQR = Q_3 - Q_1$$



For our data on North Carolina travel times, $IQR = 35 - 12 = 23$ minutes. However, *no single numerical measure of variability, such as IQR, is very useful for describing skewed distributions.* The two sides of a skewed distribution have different variability, so one number can't summarize them. That's why we give the full five-number summary. The interquartile range is mainly used as the basis for a rule of thumb for identifying suspected outliers.

The $1.5 \times IQR$ Rule for Outliers

Call an observation a suspected outlier if it falls more than $1.5 \times IQR$ above the third quartile or below the first quartile.

EXAMPLE 2.6 Using the $1.5 \times IQR$ Rule

For the North Carolina travel time data, $IQR = 23$ and

$$1.5 \times IQR = 1.5 \times 23 = 34.5$$

Any values not falling between

$$Q_1 - (1.5 \times IQR) = 12.0 - 34.5 = -22.5 \quad \text{and}$$

$$Q_3 + (1.5 \times IQR) = 35 + 34.5 = 69.5$$

are flagged as suspected outliers. Look again at the stemplot in Example 2.3 (page 48): the only suspected outlier is the longest travel time, 70 minutes. The $1.5 \times IQR$ rule suggests that the two next-longest travel times of 40 minutes are just part of the long right tail of this skewed distribution.

In a modified boxplot, which is provided by many software packages, the suspected outliers are identified in the boxplot with a special plotting symbol such as a dot (•). Comparing Figure 2.2 with Figure 2.1, we see that the largest observation from North Carolina is flagged as an outlier. The line

*This short section is optional.

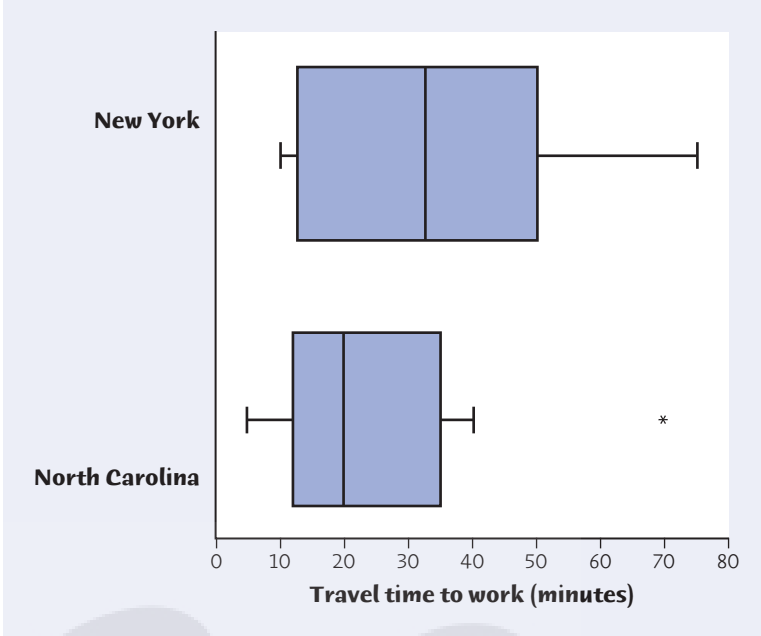
STATISTICS IN YOUR WORLD

How Much Is That House Worth?

The town of Manhattan, Kansas, is sometimes called “the Little Apple” to distinguish it from that other Manhattan, “the Big Apple.” A few years ago, a house there appeared in the county appraiser's records valued at \$200,059,000. That would be quite a house, even on Manhattan Island. As you might guess, the entry was wrong: the true value was \$59,500. But before the error was discovered, the county, the city, and the school board had based their budgets on the total appraised value of real estate, which the one outlier jacked up by 6.5%. It can pay to spot outliers before you trust your data.




FIGURE 2.2
Horizontal modified boxplots
comparing the travel times to work of
samples of workers in North Carolina
and New York.




beginning at the third quartile no longer extends to the maximum but now ends at 40, which is the largest observation from North Carolina that is not identified as an outlier. Figure 2.2 also displays the modified boxplots horizontally rather than vertically, an option available in some software packages that does not change the interpretation of the plot. Finally, the $1.5 \times IQR$ rule is not a replacement for looking at the data. It is most useful when large volumes of data are processed automatically.

APPLY YOUR KNOWLEDGE



2.8 Travel Time to Work. In Example 2.3 (page 48), there is one long travel time of 75 minutes in our sample of 20 New York travel times. Does the $1.5 \times IQR$ rule identify this travel time as a suspected outlier?  NYTRAVEL

2.9 Fuel Economy for Midsize Cars. Exercise 2.7 (page 54) gives the estimated miles per gallon (mpg) for city driving for the 189 cars classified as midsize in 2019. Are any of the larger values outliers by the $1.5 \times IQR$ rule? Although outliers can be produced by errors or incorrectly recorded observations, they are often observations that differ from the others in some particular way. In this case, the cars producing the high outliers share a common feature. What do you think that is?  MIDCARS

2.7 Measuring Variability: The Standard Deviation

The five-number summary is not the most common numerical description of a distribution. That distinction belongs to the combination of the mean to measure center and the *standard deviation* to measure variability. The standard deviation and its close relative the *variance* measure variability by looking at how far the observations are from their mean.

The Standard Deviation *s*

The **variance** s^2 of a set of observations is an average of the squares of the deviations of the observations from their mean. In symbols, the variance of n observations x_1, x_2, \dots, x_n is

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

or, more compactly,

$$s^2 = \frac{1}{n - 1} \sum (x_i - \bar{x})^2$$

The **standard deviation** s is the square root of the variance s^2 :

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

In practice, use software or your calculator to obtain the standard deviation from keyed-in data. Doing an example step-by-step will help you understand how the variance and standard deviation work, however.

EXAMPLE 2.7 Calculating the Standard Deviation

Georgia Southern University had 2786 students with regular admission in its freshman class of 2015. For each student, data are available on their SAT and ACT scores (if taken), high school GPA, and the college within the university to which they were admitted.¹⁰ In Exercise 3.49 (page 96), the full data set for the SAT Mathematics scores will be examined. Here are the first five observations from that data set:

490 580 450 570 650

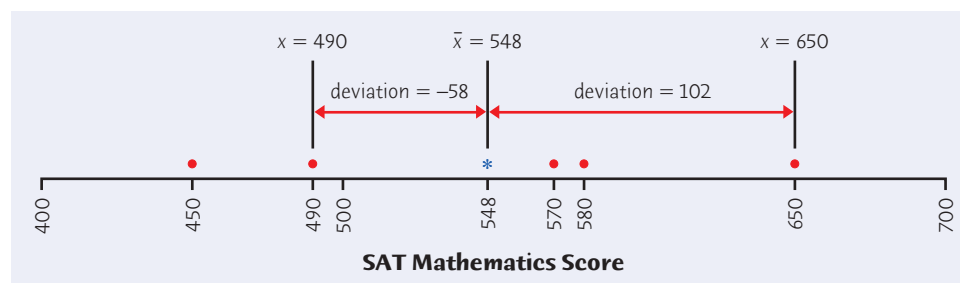
We will compute \bar{x} and s for these students. First find the mean:

$$\begin{aligned}\bar{x} &= \frac{490 + 580 + 450 + 570 + 650}{5} \\ &= \frac{2740}{5} = 548\end{aligned}$$

Figure 2.3 displays the data as points above the number line, with their mean marked by an asterisk (*). The arrows mark two of the deviations from the mean. The deviations show how variable the data are about their mean. They are the starting point for calculating the variance and the standard deviation.

Observations x_i	Deviations $x_i - \bar{x}$	Squared Deviations $(x_i - \bar{x})^2$
490	$490 - 548 = -58$	$(-58)^2 = 3,364$
580	$580 - 548 = 32$	$32^2 = 1,024$
450	$450 - 548 = -98$	$(-98)^2 = 9,604$
570	$570 - 548 = 22$	$22^2 = 484$
650	$650 - 548 = 102$	$102^2 = 10,404$
	sum = 0	sum = 24,880



**FIGURE 2.3**

SAT Mathematics scores for five students, with their mean (*) and the deviations of two observations from the mean shown, for Example 2.7.

The variance is the sum of the squared deviations divided by one less than the number of observations:

$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 = \frac{24,880}{4} = 6220$$

The standard deviation is the square root of the variance:

$$s = \sqrt{6220} = 78.87$$

Notice that the “average” in the variance s^2 divides the sum by one fewer than the number of observations, that is, $n - 1$ rather than n . The reason is that the deviations $x_i - \bar{x}$ always sum to exactly 0 so that knowing $n - 1$ of them determines the last one. Only $n - 1$ of the squared deviations can vary freely, and we average by dividing the total by $n - 1$. The number $n - 1$ is called the *degrees of freedom*¹¹ of the variance or standard deviation. Some calculators offer a choice between dividing by n and dividing by $n - 1$, so be sure to use $n - 1$.

More important than the details of hand calculation are the properties that determine the usefulness of the standard deviation:

- s measures *variability about the mean* and should be used only when the mean is chosen as the measure of center.
- s is *always zero or greater than zero*. $s = 0$ only when there is no variability. This happens only when all observations have the same value. Otherwise, $s > 0$. As the observations become more variable about their mean, s gets larger.
- s has the *same units of measurement as the original observations*. For example, if you measure weight in kilograms, both the mean \bar{x} and the standard deviation s are also in kilograms. This is one reason to prefer s to the variance s^2 , which would be in squared kilograms.
- Like the mean \bar{x} , s is *not resistant*. A few outliers can make s very large.



The use of squared deviations renders s even more sensitive than \bar{x} to a few extreme observations. For example, the standard deviation of the travel times for the 15 North Carolina workers in Example 2.1 is 16.97 minutes. (Use your calculator or software to verify this.) If we omit the high outlier, the standard deviation drops to 12.07 minutes.

If you feel that the importance of the standard deviation is not yet clear, you are right. We will see in Chapter 3 that the standard deviation is the natural measure of variability for a very important class of symmetric distributions, the Normal distributions. The usefulness of many statistical procedures is tied to distributions of particular shapes. This is certainly true of the standard deviation.

2.8 Choosing Measures of Center and Variability

We now have a choice between two descriptions of the center and variability of a distribution: the five-number summary or \bar{x} and s . Because \bar{x} and s are sensitive to extreme observations, they can be misleading when a distribution is strongly skewed or has outliers. In fact, because the two sides of a skewed distribution have different variability, no single number describes the variability well. The five-number summary, with its two quartiles and two extremes, does a better job.

Choosing a Summary

- The five-number summary is usually better than the mean and standard deviation for describing a skewed distribution or a distribution with strong outliers.
- Use \bar{x} and s only for reasonably symmetric distributions that are free of outliers.

Outliers can greatly affect the values of the mean \bar{x} and the standard deviation s , the most common measures of center and variability. Many more elaborate statistical procedures also can't be trusted when outliers are present. *Whenever you find outliers in your data, try to find an explanation for them.* Sometimes the explanation is as simple as a typing error, such as typing 10.1 as 101; if this is the case, correct the typing error. Sometimes a measuring device broke down or a subject gave a frivolous response, like the student in a class survey who claimed to study 30,000 minutes per night. (Yes, that really happened.) In all these cases, you can simply remove the outlier from your data. When outliers are “real data,” like the long travel times of some New York workers, you should choose statistical methods that are not greatly disturbed by the outliers. For example, use the five-number summary rather than \bar{x} and s to describe a distribution with extreme outliers. We will meet other examples later in the book.

Remember that a graph gives the best overall picture of a distribution. If data have been entered into a calculator or statistical program, it is very simple and quick to create several graphs to see all the different features of a distribution. Numerical measures of center and variability report specific facts about a distribution, but they do not describe its entire shape. Numerical summaries do not disclose the presence of multiple peaks or clusters, for example. Exercise 2.11 shows how misleading numerical summaries can be. *Always plot your data.*

APPLY YOUR KNOWLEDGE


2.10 \bar{x} and s by Hand. Radon is a naturally occurring gas and is the second leading cause of lung cancer in the United States.¹² It comes from the natural breakdown of uranium in the soil and enters buildings through cracks and other holes in foundations. Radon is found throughout the United States, but levels vary considerably from state to state. Several methods can reduce the levels of radon in a home, and the Environmental Protection Agency recommends using one of them if the measured level in a home is above 4 picocuries per liter. Four readings from Franklin County, Ohio, where the county average is 8.2 picocuries per liter, were 3.8, 1.9, 12.1, and 14.4.

- Find the mean step-by-step. That is, find the sum of the four observations and divide by 4.
- Find the standard deviation step-by-step. That is, find the deviation of each observation from the mean, square the deviations, and obtain the variance and the standard deviation. Example 2.7 (page 57) shows the method.



Doug Martin/Science Source

(c) Now enter the data into your calculator and use the mean and standard deviation buttons to obtain \bar{x} and s . Do the results agree with your hand calculations?

2.11 \bar{x} and s Are Not Enough. The mean \bar{x} and standard deviation s measure center and variability but are not a complete description of a distribution. Data sets with different shapes can have the same mean and standard deviation. To demonstrate this fact, use your calculator to find \bar{x} and s for these two small data sets. Then make a stemplot of each and comment on the shape of each distribution.  DATASET2

Data A:	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74
Data B:	6.58	5.76	7.71	8.84	8.47	7.04	5.25	6.89	5.56	7.91	12.50

2.12 Choose a Summary. The shape of a distribution is a rough guide to whether the mean and standard deviation are a helpful summary of center and variability. For which of the following distributions would \bar{x} and s be useful? In each case, give a reason for your decision.

- (a) Percentages of high school graduates in the states taking the SAT, Figure 1.8 (page 26)
- (b) Iowa Tests scores, Figure 1.7 (page 26)
- (c) New York travel times, Example 2.3 (page 48)

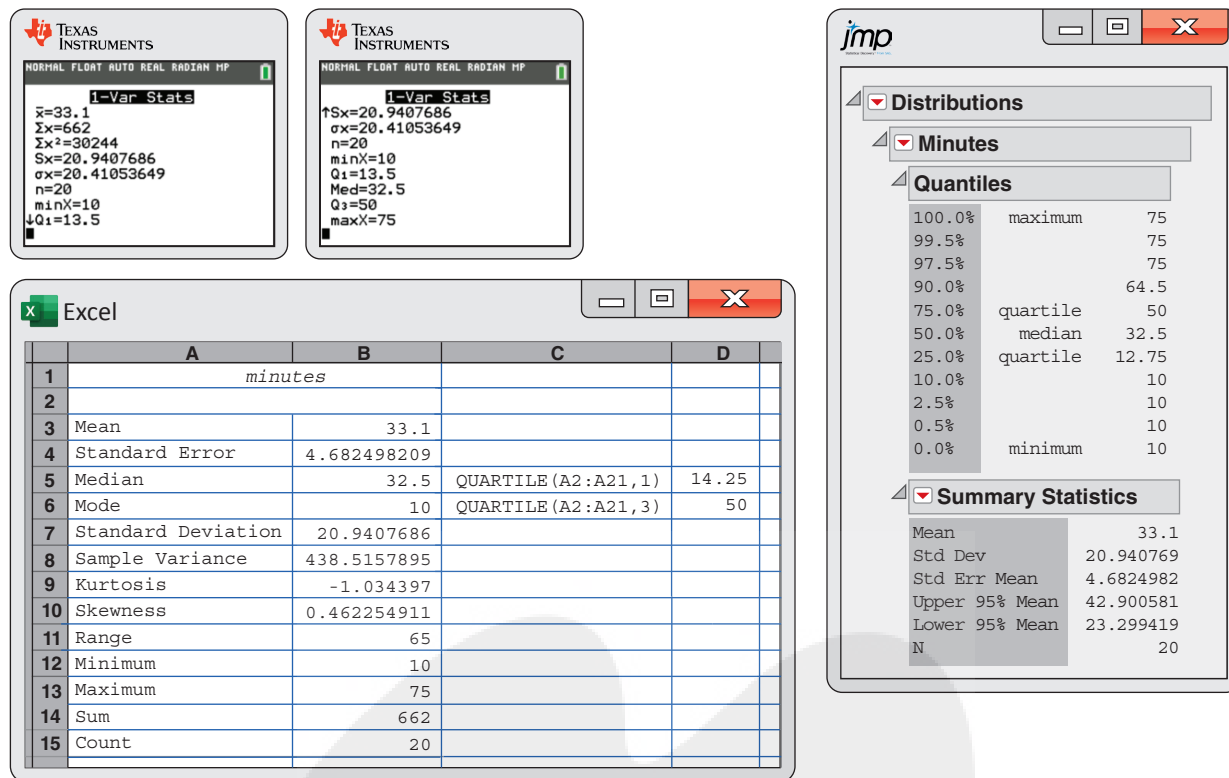
2.9 Examples of Technology

Although a calculator with “two-variable statistics” functions will do many of the basic calculations we need throughout the text, more elaborate technology is helpful. Graphing calculators and computer software will do calculations and make graphs as you command, freeing you to concentrate on choosing the right methods and interpret your results. Figure 2.4 displays outputs from three technology tools for describing the travel times to work of 20 people in New York State (Example 2.3, page 48). Can you find \bar{x} , s , and the five-number summary in each output? The big message is: *once you know what to look for, you can read output from any technological tool.*

The displays in Figure 2.4 come from a Texas Instruments graphing calculator, the Microsoft Excel spreadsheet program, and JMP statistical software. JMP allows you to choose what descriptive measures you want, whereas the descriptive measures in Excel and the calculator give some things you don’t need. Just ignore the extras. Because Excel’s “Descriptive Statistics” menu item doesn’t give the quartiles, we used the spreadsheet’s separate quartile function to get Q_1 and Q_3 .

EXAMPLE 2.8 What Is the Third Quartile?

In Example 2.5, we saw that the quartiles of the New York travel times are $Q_1 = 13.5$ and $Q_3 = 50$. Look at the output displays in Figure 2.4. The calculator agrees with our work, while Excel says $Q_1 = 14.25$ and JMP says that $Q_1 = 12.75$. What happened? *There are several rules for finding the quartiles. Some calculators and software use rules that give results different from ours for some sets of data.* This is true of JMP and Excel. Results from the various rules are generally close to each other, so the differences are not important in practice. Our rule is the simplest for hand calculation.

**FIGURE 2.4**

Output from a graphing calculator, a spreadsheet program, and a statistical software package describing the data on travel times to work in New York State.

2.10 Organizing a Statistical Problem

Most of our examples and exercises have aimed to help you learn basic tools (graphs and calculations) for describing and comparing distributions. You have also learned principles that guide use of these tools, such as “start with a graph” and “look for the overall pattern and striking deviations from the pattern.” The data you work with are not just numbers; they describe specific settings such as water depth in the Everglades or travel time to work. Because data come from a specific setting, the final step in examining data is *coming to a conclusion for that setting*. Water depth in the Everglades has a yearly cycle that reflects Florida’s wet and dry seasons. Travel times to work are generally longer in New York than in North Carolina.

Let’s return to the on-time high school graduation rates discussed in Example 1.4 (page 20). We know from the example that the on-time graduation rates vary from 71.1% in New Mexico to 91% in Iowa, with a median of 86%. State graduation rates are related to many factors, and in a statistical problem, we often try to explain the differences or variation in a variable such as graduation rate by some of these factors. For example, do states with lower household incomes tend to have lower high school graduation rates? Or, do the states in some regions of the country tend to have lower high school graduation rates than the states in other regions?

As you learn more statistical tools and principles, you will face more complex statistical problems. Although no framework accommodates all the varied issues that arise in applying statistics to real settings, we find the following four-step thought process gives useful guidance. In particular, the first and last steps emphasize that

statistical problems are tied to specific real-world settings and therefore involve more than doing calculations and making graphs.

Organizing a Statistical Problem: A Four-Step Process

STATE: What is the practical question, in the context of the real-world setting?

PLAN: What specific statistical operations does this problem call for?

SOLVE: Make the graphs and carry out the calculations needed for this problem.

CONCLUDE: Give your practical conclusion in the setting of the real-world problem.

To help you master the basics, many exercises will continue to tell you what to do—make a histogram, find the five-number summary, and so on. Real statistical problems don't come with detailed instructions. From now on, especially in the later chapters of the book, you will meet some exercises that are more realistic. Use the four-step process as a guide to solving and reporting these problems. They are marked with the four-step icon, as the following example illustrates.

EXAMPLE 2.9 Comparing Graduation Rates



STATE: Federal law requires all states in the United States to use a common computation of on-time high school graduation rates beginning with the 2010–11 school year. Previously, states chose one of several computation methods that gave answers that could differ by more than 10%. This common computation allows for meaningful comparison of graduation rates between the states.

We know from Table 1.1 (page 21), that the on-time high school graduation rates in the 2016–17 school year varied from 71.1% in New Mexico to 91% in Iowa. The U.S. Census Bureau divides the 50 states and the District of Columbia into four geographical regions: the Northeast (NE), Midwest (MW), South (S), and West (W). The region for each state is included in Table 1.1. Do the states in the four regions of the country display distinct distributions of graduation rates? How do the mean graduation rates of the states in each of these regions compare?

PLAN: Use graphs and numerical descriptions to describe and compare the distributions of on-time high school graduation rates of the states in the four regions of the United States.

SOLVE: We might use boxplots to compare the distributions, but stemplots preserve more detail and work well for data sets of these sizes. Figure 2.5 displays the stemplots with the stems lined up for easy comparison. The stems have been split to better display the distributions, and the data have been rounded to the nearest percentage (with no decimal places). The stemplots overlap, and some care is needed when comparing the four stemplots because the sample sizes differ, with some stemplots having more leaves than others. The states in the Northeast and Midwest have distributions that are similar to each other. The South, with the most observations, has one low observation corresponding to the District of Columbia that stands apart from the others and some skewness to the left. With little skewness and no serious outliers, we report \bar{x} and s as our summary measures of center and variability of the distribution of the on-time graduation rates of the states in each region. Because the District of Columbia is not a state, although often included with state data, we have reported summary statistics for the South with and without this observation.

Region	<i>n</i>	Mean	Standard Deviation
Midwest	12	86.03	3.12
Northeast	9	87.12	2.70
South (including DC)	17	85.14	4.72
South (excluding DC)	16	85.89	3.69
West	13	80.5	4.27

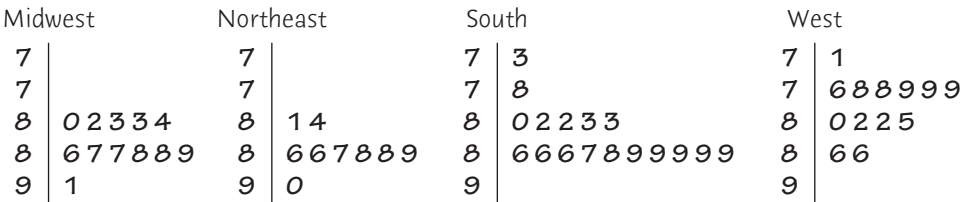



FIGURE 2.5
Stemplots comparing the distributions of graduation rates for the four census regions from Table 1.1, for Example 2.9.

CONCLUDE: The table of summary statistics and the stemplots lead to similar conclusions. The states in the Midwest and Northeast are most similar to each other, with the South, excluding the District of Columbia, having a slightly lower mean and higher standard deviation. The states in the West have a lower mean graduation rate than the other three regions, with a standard deviation similar to that of the South but higher than those of the Midwest or Northeast.

It is important to remember that the individuals in Example 2.9 are the states. For example, the mean of 87.12 is the mean of the on-time graduation rates for the nine Northeastern states, and the standard deviation tells us how much these state rates vary about this mean. However, the mean of these nine states is *not* the same as the graduation rate for all high school students in the Northeast, unless the states have the same number of high school graduates. The graduation rate for all high school students in the Northeast would be a *weighted* average of the state rates, with the larger states receiving more weight. For example, because New York is the most populous state in the Northeast and also has the lowest graduation rate, we would expect the graduation rate of all high school students in the Northeast to be lower than 87.12 because New York would pull down the overall graduation rate. See Exercise 2.37 (page 68) for a similar example.

APPLY YOUR KNOWLEDGE

2.13 Logging in the Rain Forest. “Conservationists have despaired over destruction of tropical rain forest by logging, clearing, and burning.” These words begin a report on a statistical study of the effects of logging in Borneo.¹³ Charles Cannon of Duke University and his coworkers compared forest plots that had never been logged (Group 1) with similar plots nearby that had been logged one year earlier (Group 2) and eight years earlier (Group 3). Each plot was 0.1 hectare in area. Here are the counts of trees for plots in each group:  LOGGING

Group 1:	27	22	29	21	19	33	16	20	24	27	28	19
Group 2:	12	12	15	9	20	18	17	14	14	2	17	19
Group 3:	18	4	22	15	18	19	22	12	12			

To what extent has logging affected the count of trees? Follow the four-step process in reporting your work.



AustralianCamera/Shutterstock



2.14 Worldwide Child Mortality. Although child mortality rates worldwide have dropped by more than 50% since 1990, in 2017 it was still the case that 16,000 children under five years old died each day. The mortality rates for children under five varied from 2.1 per 1000 in Iceland to 127.2 per 1000 in Somalia. In Exercise 1.36 (page 40), you were asked to draw a histogram of these data. In this exercise, you will explore the relationship between child mortality and a measure of a country's economic wealth. One measure used by the World Bank is the gross national income (GNI) per capita, the dollar value of a country's final income in a year divided by its population. It reflects the average income of a country's citizens, and the World Bank uses GNI per capita to classify countries into low-income, lower-middle-income, upper-middle-income, and high-income economies. Although the data set includes 214 countries, the child mortality rates of 21 countries are not available in the World Health Organization database. Because the data set is too large to print here, we give the data for the first five countries:¹⁴

Country	Child Mortality Rate (per 1000)	Economy Classification
Aruba	—	High
Afghanistan	67.9	High
Angola	81.1	Low
Albania	8.8	Upper-middle
Andorra	3.3	Upper-middle

Give a full description of the distribution of child mortality rates for the countries in each of the four economic classifications and identify any high outliers. Compare the four groups. Does the economic classification used by the World Bank do a good job of explaining the differences in child mortality rates among the countries?

CHAPTER 2 SUMMARY

- A numerical summary of a distribution should report at least its center and its variability.
- The **mean** \bar{x} and the **median** M describe the center of a distribution in different ways. The mean is the arithmetic average of the observations, and the median is the midpoint of the values.
- When you use the median to indicate the center of the distribution, describe its variability by giving the **quartiles**. The **first quartile**, Q_1 , has one-fourth of the observations below it, and the **third quartile**, Q_3 , has three-fourths of the observations below it.
- The **five-number summary** consisting of the median, the quartiles, and the smallest and largest individual observations provides a quick overall description of a distribution. The median describes the center, and the quartiles and extremes show the variability.
- **Boxplots** based on the five-number summary are useful for comparing several distributions. The box spans the quartiles and shows the variability of the central half of the distribution. The median is marked within the box. Lines extend from the box to the extremes and show the full variability of the data.
- The **variance** s^2 and especially its square root, the **standard deviation** s , are common measures of variability about the mean as center. The standard deviation s is zero when there is no variability and gets larger as the variability increases.

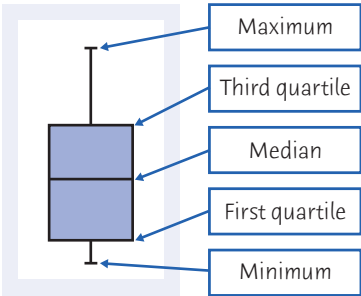






FIGURE 2.6
Boxplot showing maximum, third quartile, median, first quartile, and minimum.

- A **resistant measure** of any aspect of a distribution is relatively unaffected by changes in the numerical value of a small proportion of the total number of observations, no matter how large these changes are. The median and quartiles are resistant, but the mean and the standard deviation are not.
- The mean and standard deviation are good descriptions for symmetric distributions without outliers. They are most useful for the Normal distributions introduced in the next chapter. The five-number summary is a better description for skewed distributions.
- Numerical summaries do not fully describe the shape of a distribution. Always plot your data.
- A statistical problem has a real-world setting. You can organize many problems by using the following four steps: *state*, *plan*, *solve*, and *conclude*.


CHECK YOUR SKILLS

- 2.15 The 2019–20 roster of the New England Patriots, winners of the 2019 NFL Super Bowl, included 10 defensive linemen. The weights in pounds of the 10 defensive linemen were  LINEMEN
275 300 300 315 345 260 250 275 280 250
The mean of these data is
(a) 277.50. (b) 285.00. (c) 300.25.
- 2.16 The median of the data in Exercise 2.15 is  LINEMEN
(a) 277.50. (b) 285.00. (c) 300.25.
- 2.17 The first quartile of the data in Exercise 2.15 is  LINEMEN
(a) 260.00. (b) 300.00. (c) 303.75.
- 2.18 If a distribution is skewed to the left,
(a) the mean is less than the median.
(b) the mean and median are equal.
(c) the mean is greater than the median.
- 2.19 What percentage of the observations in a distribution are greater than the first quartile?
(a) 25% (b) 50% (c) 75%
- 2.20 To make a boxplot of a distribution, you must know
(a) all the individual observations.
(b) the mean and the standard deviation.
(c) the five-number summary.
- 2.21 The standard deviation of the 10 weights in Exercise 2.15 (use your calculator) is about  LINEMEN
(a) 28.72. (b) 30.28. (c) 46.25.
- 2.22 What are all the values that a standard deviation s can possibly take?
(a) $0 \leq s$ (b) $0 \leq s \leq 1$ (c) $-1 \leq s \leq 1$
- 2.23 The correct units for the standard deviation in Exercise 2.21 are
(a) no units—it's just a number.
(b) pounds.
(c) pounds squared.
- 2.24 Which of the following is most affected if an extreme high outlier is added to the data?
(a) The median
(b) The mean
(c) The first quartile

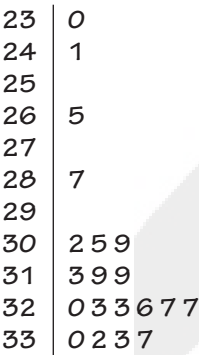
CHAPTER 2 EXERCISES

- 2.25 **Incomes of College Grads.** According to the U.S. Census Bureau's Current Population Survey, the mean and median 2018 income of people aged 25–34 years who had a bachelor's degree but no higher degree were \$50,350 and \$60,178.¹⁵ Which of these numbers is the mean, and which is the median? Explain your reasoning.
- 2.26 **Household Assets.** Once every three years, the Board of Governors of the Federal Reserve System collects data on household assets and liabilities through the Survey of Consumer Finances (SCF).¹⁶ Here are some results from the 2016 survey.
- (a) Transaction accounts, which include checking, savings, and money market accounts, are the most commonly held type of financial asset. The mean value of transaction accounts per household for those holding transaction accounts was \$40,200, and the median value was \$4,500. What explains the differences between the two measures of center?
- (b) The median value of cash value life insurance per household was \$0. What does a median of \$0 say about the percentage of households with cash value life insurance?


2.27 **University Endowments.** The National Association of College and University Business Officers collects data on college endowments. In 2018, its report included the endowment values of 809 colleges and universities in the United States and Canada. When the endowment values are arranged in order, what are the locations of the median and the quartiles in this ordered list?

2.28 **Pulling Apart Wood.** Exercise 1.46 (page 44) gives the breaking strengths in pounds of 20 pieces of Douglas fir.  WOOD

- (a) Give the five-number summary of the distribution of breaking strengths.
- (b) Here is a stemplot of the data rounded to the nearest hundred pounds. The stems are thousands of pounds, and the leaves are hundreds of pounds.




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

2.29 **Comparing Graduation Rates.** An alternative presentation to compare the graduation rates in Table 1.1 (page 21) by region of the country reports five-number summaries and uses boxplots to display the distributions. Calculate the five-number summaries and make the boxplots. Do the boxplots fail to reveal any important information visible in the stemplots of Figure 2.5 (page 63)? Which plots make it simpler to compare the regions? Why?  GRADRATE

2.30 **How Much Fruit Do Adolescent Girls Eat?** Figure 1.17 (page 37) is a histogram of the number of servings of fruit per day claimed by 74 17-year-old girls.

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

(c) In general, you cannot find the exact values of the median, quartiles, or mean from the histogram. What is special about the histogram of the number of servings of fruit that allows you to do this?

2.31 **Guinea Pig Survival Times.** Here are the survival times, in days, of 72 guinea pigs after they were injected with infectious bacteria in a medical experiment.¹⁷ Survival times, whether of machines under stress or cancer patients after treatment, usually have distributions that are skewed to the right.  GUINPIGS

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

- (a) Graph the distribution and describe its main features. Does it show the expected right-skew?
- (b) Which numerical summary would you choose for these data? Calculate your chosen summary. How does it reflect the skewness of the distribution?

2.32 **Maternal Age at Childbirth.** How old are women when they have their first child? Here is the distribution of the age of the mother for all firstborn children in the United States in 2017.¹⁸

Age	Count	Age	Count
10–14 years	1,892	30–34 years	329,623
15–19 years	162,536	35–39 years	124,637
20–24 years	395,927	40–44 years	24,049
25–29 years	417,162	45–49 years	2,377


The number of first-born children to mothers under 10 or over 50 years of age represent a negligible percentage of all first births and are not included in the table.



Tom Merton/Getty Images

- (a) For comparison with other years and with other countries, we prefer a histogram of the percentages in each age class rather than the counts. Explain why.
- (b) How many babies were there?

- (c) Make a histogram of the distribution, using percentages on the vertical scale. Using this histogram, describe the distribution of the age at which women have their first child.
- (d) What are the locations of the median and quartiles in the ordered list of all maternal ages? In which age classes do the median and quartiles fall?

2.33 More on Nintendo and Laparoscopic Surgery. In Exercise 1.38 (page 41), you examined the improvement in times to complete a virtual gall bladder removal for those with and without four weeks of Nintendo Wii™ training. The most common methods for formal comparison of two groups use \bar{x} and s to summarize the data.  NINTENDO

- (a) What kinds of distributions are best summarized by \bar{x} and s ? Do you think these summary measures are appropriate in this case?
- (b) In the control group, one subject improved his/her time by 229 seconds. How much does removing this observation change \bar{x} and s for the control group? You will need to compute \bar{x} and s for the control group, both with and without the high outlier.
- (c) Compute the median for the control group with and without the high outlier. What does this show about the resistance of the median and \bar{x} ?

2.34 Making Resistance Visible. In the *Mean and Median* applet, place three observations on the line by clicking below it: two close together near the center of the line and one somewhat to the right of these two.



2.35



Behavior of the Median. Place five observations on the line in the *Mean and Median* applet by clicking below it.

- (a) Add one additional observation *without changing the median*. Where is your new point?
- (b) Use the applet to convince yourself that when you add yet another observation (there are now seven in all), the median does not change, no matter where you put the seventh point. Explain why this must be true.

2.36

Never on Sunday: Also in Canada? Exercise 1.5 (page 19) gives the number of births in the United States on each day of the week during an entire year. The boxplots in Figure 2.7 are based on more detailed data from Toronto, Canada: the number of births on each of the 365 days in a year, grouped by day of the week.¹⁹ Based on these plots, compare the day-of-the-week distributions using shape, center, and variability. Summarize your findings.

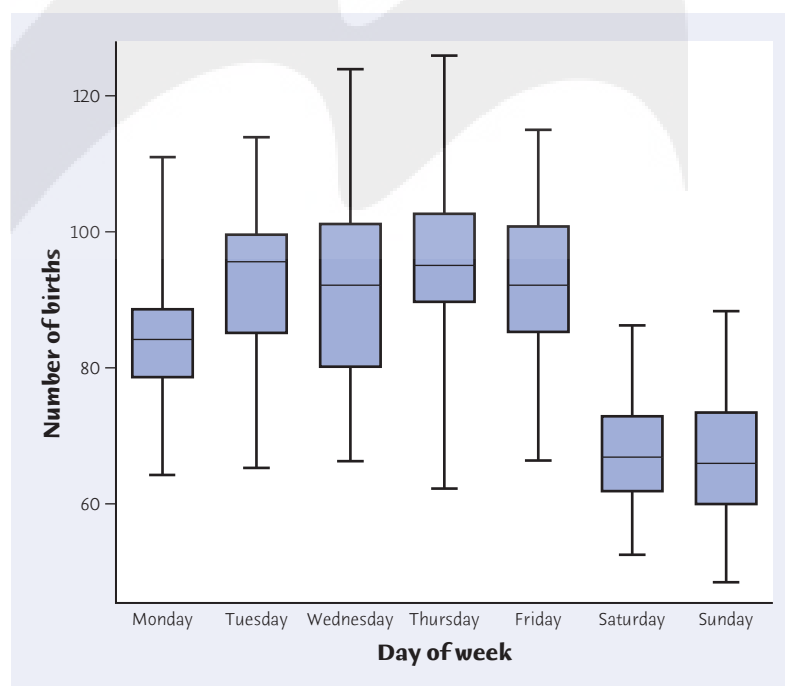



FIGURE 2.7

Boxplots of the distributions of numbers of births in Toronto, Canada, on each day of the week during a year, for Exercise 2.36.

2.37 **Thinking about Means.** Table 1.2 (page 23) gives the percentage of minority residents in each of the states. For the nation as a whole, 42.8% of residents are minorities. Find the mean of the 51 entries in Table 1.2. It is *not* 42.8%. Explain carefully why this happens. (*Hint:* The states with the largest populations are California, Texas, New York, and Florida. Look at their entries in Table 1.2.)  MINORITY

2.38 **Thinking about Means and Medians.** In 2018, approximately 2.1% of hourly rate workers were being paid at the federal minimum wage level or lower. Would federal legislation to increase the minimum wage have a greater effect on the mean or the median income of *all* workers? Explain your answer.

2.39 **A Standard Deviation Contest.** You are to choose four numbers from the whole numbers 0 to 10, with repeats allowed.


- Choose four numbers that have the smallest possible standard deviation.
- Choose four numbers that have the largest possible standard deviation.
- Is more than one choice possible in either part (a) or (b)? Explain.

2.40 **You Create the Data.** Create a set of seven numbers (repeats allowed) that have the five-number summary

Minimum = 4 $Q_1 = 8$ $M = 12$ $Q_3 = 15$ Maximum = 19

There is more than one set of seven numbers with this five-number summary. What must be true about the seven numbers to have this five-number summary?

2.41 **You Create the Data.** Give an example of a small set of data for which the mean is greater than the third quartile.

2.42 **Adolescent Obesity.** Adolescent obesity is a serious health risk affecting more than 5 million young people in the United States alone. Laparoscopic adjustable gastric banding has the potential to provide a safe and effective treatment. Fifty adolescents between 14 and 18 years old with a body mass index (BMI) higher than 35 were recruited from the Melbourne, Australia, community for the study.²⁰ Twenty-five were randomly selected to undergo gastric banding, and the remaining 25 were assigned to a supervised lifestyle intervention program involving diet, exercise, and behavior modification. All subjects were followed for two years. Here are the weight losses, in kilograms, for the subjects who completed the study:  GASTRIC

Gastric Banding					
35.6	81.4	57.6	32.8	31.0	37.6
36.5	-5.4	27.9	49.0	64.8	39.0
43.0	33.9	29.7	20.2	15.2	41.7
53.4	13.4	24.8	19.4	32.3	22.0


Lifestyle Intervention					
6.0	2.0	-3.0	20.6	11.6	15.5
-17.0	1.4	4.0	-4.6	15.8	34.6
6.0	-3.1	-4.3	-16.7	-1.8	-12.8

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

Exercises 2.43 through 2.49 ask you to analyze data without having the details outlined for you. The exercise statements give you the **state** step of the four-step process. In your work, follow the **plan**, **solve**, and **conclude** steps as illustrated in Example 2.9 (page 62).



2.43 **Protective Equipment and Risk Taking.** Studies have shown that people who are using safety equipment when engaging in an activity tend to take increased risks. Will risk taking increase when people are not aware that they are wearing protective equipment and are engaged in an activity that cannot be made safer by this equipment? Participants in the study were falsely told they were taking part in an eye-tracking experiment for which they needed to wear an eye-tracking device. Eighty subjects were divided at random into two groups of 40 each, with one group wearing the tracking device mounted on a baseball cap and the other group wearing it mounted on a bicycle helmet. Subjects were told that the helmet or cap was just being used to mount the eye tracker. All subjects watched an animated balloon on a video screen and pressed a button to inflate it. The balloon was programmed to burst at a random point, but until that point, each press of the button inflated the balloon further and increased the amount of fictional


currency a subject would earn. Subjects were free to stop pumping at any point and keep their earnings, knowing that if the balloon burst, they would lose all earnings for that round. The score was the average number of pumps on the trials, with lower scores corresponding to less risk taking and more conservative play. Here are the first 10 observations from each group:²¹  HELMET




Tim Gamble and Ian Walker, "Wearing a bicycle helmet can increase risk taking and sensation seeking in adults," *Psychological Science*, 27 (2016), pp. 289–294: <https://doi.org/10.1177/0956797615620784> (<http://www.creativecommons.org/licenses/by/3.0/>).

Helmet:	3.67	36.50	29.28	30.50	24.08
	32.10	50.67	26.26	41.05	20.56
Baseball Cap:	29.38	42.50	41.57	47.77	32.45
	30.65	7.04	2.68	22.04	25.86


Compare the distributions for the two groups. How is wearing of a helmet related to the measure of risk behavior?

2.44  **Athletes' Salaries.** The Montreal Canadiens were founded in 1909 and are the longest continuously operating professional ice hockey team. The team has won 24 Stanley Cups, making them one of the most successful professional sports teams of the traditional four major sports of Canada and the United States. Table 2.1 gives the salaries of the 2019–20 roster prior

to the start of the 2019–20 season.²² Provide the team owner with a full description of the distribution of salaries and a brief summary of its most important features.  HOCKEY

2.45



Returns on Stocks. How well have stocks done over the past generation? The Wilshire 5000 index describes the average performance of all U.S. stocks. The average is weighted by the total market value of each company's stock, so think of the index as measuring the performance of the average investor. Shown below are the percentage returns on the Wilshire 5000 index for the years 1971–2018: What can you say about the distribution of yearly returns on stocks?²³  WILSHIRE



Year	Return	Year	Return	Year	Return
1971	17.68	1987	2.27	2003	31.64
1972	17.98	1988	17.94	2004	12.62
1973	−18.52	1989	29.17	2005	6.32
1974	−28.39	1990	−6.18	2006	15.88
1975	38.47	1991	34.20	2007	5.73
1976	26.59	1992	8.97	2008	−37.34
1977	−2.64	1993	11.28	2009	29.42
1978	9.27	1994	−0.06	2010	17.87
1979	25.56	1995	36.45	2011	0.59
1980	33.67	1996	21.21	2012	16.12
1981	−3.75	1997	31.29	2013	34.02
1982	18.71	1998	23.43	2014	12.07
1983	23.47	1999	23.56	2015	−0.24
1984	3.05	2000	−10.89	2016	13.04
1985	32.56	2001	−10.97	2017	21.00
1986	16.09	2002	−20.86	2018	−5.29


TABLE 2.1 Salaries for the 2019–20 Montreal Canadiens


Player	Salary	Player	Salary	Player	Salary
Carey Price	\$15,000,000	Phillip Danault	\$3,000,000	Christian Folin	\$800,000
Shea Weber	\$6,000,000	Brett Kulak	\$1,950,000	Victor Mete	\$750,000
Jonathan Drouin	\$5,500,000	Dale Weise	\$1,750,000	Charlie Lindgren	\$750,000
Tomas Tatar	\$4,981,132	Jordan Weal	\$1,300,000		
Karl Alzner	\$4,625,000	Matthew Peca	\$1,300,000		
Paul Byron	\$4,000,000	Nate Thompson	\$1,000,000		
Jeff Petry	\$4,000,000	Nicolas Deslauriers	\$950,000		
Brendan Gallagher	\$4,000,000	Jesper Kotkaniemi	\$925,000		
Andrew Shaw	\$3,250,000	Ryan Poehling	\$925,000		
Max Domi	\$3,150,000	Noah Juulsen	\$832,000		

TABLE 2.2 Amount spent (euros) by customers in a restaurant when exposed to odors

No Odor									
15.9	18.5	15.9	18.5	18.5	21.9	15.9	15.9	15.9	15.9
15.9	18.5	18.5	18.5	20.5	18.5	18.5	15.9	15.9	15.9
18.5	18.5	15.9	18.5	15.9	18.5	15.9	25.5	12.9	15.9
Lemon Odor									
18.5	15.9	18.5	18.5	18.5	15.9	18.5	15.9	18.5	18.5
15.9	18.5	21.5	15.9	21.9	15.9	18.5	18.5	18.5	18.5
25.9	15.9	15.9	15.9	18.5	18.5	18.5	18.5		
Lavender Odor									
21.9	18.5	22.3	21.9	18.5	24.9	18.5	22.5	21.5	21.9
21.5	18.5	25.5	18.5	18.5	21.9	18.5	18.5	24.9	21.9
25.9	21.9	18.5	18.5	22.8	18.5	21.9	20.7	21.9	22.5


2.46  **Do Good Smells Bring Good Business?** Businesses know that customers often respond to background music. Do they also respond to odors? Nicolas Guéguen and his colleagues studied this question in a small pizza restaurant in France on Saturday evenings in May. On one evening, a relaxing lavender odor was spread through the restaurant; on another evening, a stimulating lemon odor; a third evening served as a control, with no odor. Table 2.2 shows the amounts (in euros) that customers spent on each of these evenings.²⁴ Compare the three distributions. Were both odors associated with increased customer spending?  **ODORS**

2.47  **Policy Justification: Pragmatic vs. Moral.** How does a leader's justification of his/her organization's policy affect support for the policy? A study compared a moral, pragmatic, and ambiguous justification for three policy proposals: a politician's plan to fund a retirement planning agency, a state governor's plan to repave state highways, and a president's plan to outlaw child labor in a developing country. For example, for the retirement agency proposal, the moral justification was the importance of retirees "to live with dignity and comfort," the pragmatic was "to not drain public funds," and the ambiguous was "to have sufficient funds." Three hundred seventy-four volunteer subjects were assigned at random to read all three proposals: 122 subjects read the three proposals with a moral justification, 126 subjects read the three proposals with a pragmatic justification, and 126 subjects read the three proposals with an ambiguous justification. Several questions measuring support for each policy proposal were answered by each subject to create a support score for each proposal, and their scores

for the three proposals were then averaged to create an index of policy support for each subject, higher values indicating greater support.²⁵ Here are the first five observations:  **JUSTIFY**

Justification:	Pragmatic	Ambiguous	Pragmatic	Moral	Ambiguous
Policy support index:	5	7	4.75	7	5.75


The first individual read the proposals with a pragmatic justification, with a policy support index of 5, the second with an ambiguous justification and a policy support index of 7, and so forth. Compare the three distributions. How does the support index vary with the type of justification?

2.48  **Does Playing Video Games Improve Surgical Skill?** In

laparoscopic surgery, a video camera and several thin instruments are inserted into the patient's abdominal cavity. The surgeon uses the image from the video camera positioned inside the patient's body to perform the procedure by manipulating the instruments that have been inserted. The Top Gun Laparoscopic Skills and Suturing Program was developed to help surgeons develop the skill set necessary for laparoscopic surgery. Because of the similarity in many of the skills involved in video games and laparoscopic surgery, it was hypothesized that surgeons with greater prior video game experience might acquire the skills required in



yacobehtuk/Getty Images


laparoscopic surgery more easily. Thirty-three surgeons participated in the study and were classified into the three categories—never used, under three hours per day, and more than three hours per day—depending on the number of hours they played video games at the height of their video game use. They also performed Top Gun drills and received a score based on the time to complete the drill and the number of errors made, with lower scores indicating better performance. Here are the Top Gun scores and video game categories for the 33 participants:²⁶  TOPGUN

Never played:	9379	8302	5489	5334	4605	4789	9185	7216	9930
	4828	5655	4623	7778	8837	5947			
Under three hours:	5540	6259	5163	6149	4398	3968	7367	4217	5716
Three or more hours:	7288	4010	4859	4432	4845	5394	2703	5797	3758

Compare the distributions for the three groups. How is prior video game experience related to Top Gun scores?

2.49



Cholesterol Levels and Age. The National Health and Nutrition Examination Survey (NHANES) is a unique survey that combines interviews and physical examinations.²⁷ It includes basic demographic information; questions about topics such as diet, physical activity, and prescription medications; and results of a physical examination measuring a variety of variables, including blood pressure and cholesterol levels. The program began in the early 1960s, and the survey currently examines a nationally representative sample of about 5000 persons each year. You will work with the total cholesterol measurements (mg/dL) obtained from participants in the survey in 2009–2010.  CHOLEST


To examine changes in cholesterol with age, we consider only the 3044 participants between 20 and 50 years of age and have classified them into the three age categories: 20s, 30s, and 40s. The full data set is too large to print here, but here are the first 10 individuals:

Age category:	30s	20s	20s	40s	30s	40s	20s	30s	30s	20s
Total cholesterol:	135	160	299	197	196	202	175	216	181	149

The first individual is in the 30s with total cholesterol of 135, the second in the 20s with total cholesterol of 160, and so forth.

- Use graphical and numerical summaries to compare the three distributions. How does cholesterol change with age?
- The ideal range of total cholesterol is below 200 mg/dL. For individuals with elevated cholesterol levels, prescription drugs are often recommended to reduce levels. Among the 3044 participants between 20 and 50 years of age, 4 individuals in their 20s, 24 individuals in their 30s, and 117 individuals in their 40s were taking prescription medications to reduce their cholesterol levels. How do you think your comparison of the distribution would be changed if none of the individuals were taking medication? Explain.

Exercises 2.50 through 2.53 make use of the optional material on the $1.5 \times IQR$ rule for suspected outliers.

2.50 The Changing Face of America. Figure 1.10 (page 29) gives a stemplot of the percentage of minority residents aged 18–34 in each of the 50 states and the District of Columbia. These data are given in Table 1.2 (page 23).  MINORITY

- Give the five-number summary of this distribution.
- Although there do not appear to be any outliers in Figure 1.10, when you split the stems for the data in Exercise 1.10, Texas, California, New Mexico, and Hawaii are separated from the remaining states. Are these four states outliers or just the largest observations in a strongly skewed distribution? What does the $1.5 \times IQR$ rule say?

2.51


Shared Pain and Bonding. In Exercise 2.6, you should have noticed some low outliers in the pain group.



- Compute the mean and the median of the bonding scores for the pain group, both with and without the two smallest scores. Do they have more of an effect on the mean or the median? Explain why.
- Does the $1.5 \times IQR$ rule identify these two low bonding scores as suspected outliers?
- Unusual observations are not necessarily mistakes. Suppose a small percentage of subjects would experience little bonding regardless of whether they were in the no-pain group or the pain group. Explain how the randomization of the students to the two groups could have led to these “outliers.”

2.52

The Fortune Global 500. The *Fortune* Global 500, also known as the Global 500, is an annual ranking by *Fortune* magazine of the top 500 corporations worldwide as measured by revenue. In total, the Global 500

generated \$32.7 trillion in revenues in 2018. Table 2.3 provides a list of the 30 companies with the highest revenues (in billions of dollars) in 2018.²⁸ A stemplot or histogram shows that the distribution is strongly skewed to the right.  GLOBE500

- (a) Give the five-number summary. Explain why this summary suggests that the distribution is right-skewed.
- (b) Which companies are outliers according to the $1.5 \times IQR$ rule? Make a stemplot of the data. Do you agree with the rule's suggestions about which companies are and are not outliers?

- (c) If you consider *all* 500 companies, the 30 companies in Table 2.3 each represent a high outlier among all Global 500 companies. Is there a common feature shared by many of the 30 companies in the table? What proportion of the total of the Global 500 revenues is accounted for by these 30 companies?


2.53 Cholesterol for People in Their 20s. Exercise 2.49 contains the cholesterol levels of individuals in their 20s from the NHANES survey in 2009–10. The cholesterol levels are right-skewed, with a few large cholesterol levels. Which cholesterol levels are suspected outliers by the $1.5 \times IQR$ rule?  CHOLES20

TABLE 2.3 Revenues for the top Global 500 companies in 2018			
Company Name	Revenues (\$b)	Company Name	Revenues (\$b)
Wal-Mart Stores	514.4	Glencore	219.8
Sinopec Group	414.6	McKesson	214.3
Royal Dutch Shell	396.6	Daimler	197.5
China National Petroleum	393.0	Total	184.1
State Grid	387.1	China State Construction Engineering	181.5
Saudi Aramco	355.9	Trafigura Group	180.7
BP	303.7	Hon Hai Precision Industry	175.6
Exxon Mobil	290.2	EXOR Group	175.0
Volkswagen	278.3	AT&T	170.8
Toyota Motor	272.6	Industrial & Commercial Bank of China	169.0
Apple	265.6	AmerisourceBergen	167.9
Berkshire Hathaway	247.8	Chevron	166.3
Amazon.com	232.9	Ping An Insurance	163.6
UnitedHealth Group	226.2	Ford Motor	160.3
Samsung Electronics	221.6	China Construction Bank	151.1