Mark Cornelison/Photoshot Holdings Ltd.

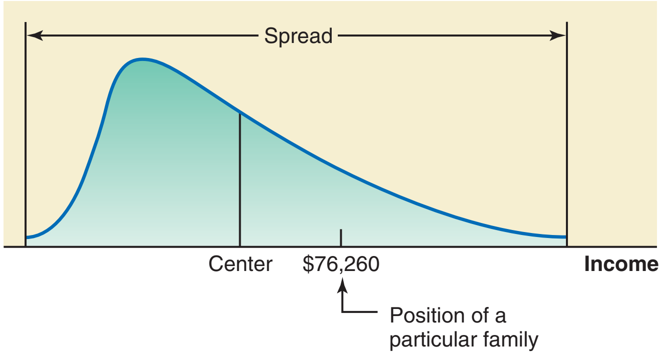
CHAPTER 3

Numerical Descriptive Measures

Is the *average* over? You may ask, “How can the *average* be over?” Well, it seems like a strange question unless you are Thomas L. Friedman, a columnist for *The New York Times*. You may ask, “Why does Mr. Friedman think that the *average* is over?” Read his article, which appears as Case Study 3–2.

In Chapter 2 we discussed how to summarize data using different methods and to display data using graphs. Graphs are one important component of statistics; however, it is also important to numerically describe the main characteristics of a data set. The numerical summary measures, such as the ones that identify the center and spread of a distribution, identify many important features of a distribution. For example, the techniques learned in Chapter 2 can help us graph data on family incomes. However, if we want to know the income of a “typical” family (given by the center of the distribution), the spread of the distribution of incomes, or the relative position of a family with a particular income, the numerical summary measures can provide more detailed information (see Figure 3.1). The measures that we discuss in this chapter include measures of (1) central tendency, (2) dispersion (or spread), and (3) position.

Figure 3.1



3.1 Measures of Central Tendency for Ungrouped Data

We often represent a data set by numerical summary measures, usually called the *typical values.* A **measure of central tendency** gives the center of a histogram or a frequency distribution curve. This section discusses three different measures of central tendency: the mean, the median, and the mode; however, a few other measures of central tendency, such as the trimmed mean, the weighted mean, and the geometric mean, are explained in exercises following this section. We will learn how to calculate each of these measures for ungrouped data. Recall from Chapter 2 that the data that give information on each member of the population or sample individually are called *ungrouped data*, whereas *grouped data* are presented in the form of a frequency distribution table.

3.1.1 Mean

The **mean**, also called the *arithmetic mean*, is the most frequently used measure of central tendency. This book will use the words *mean* and *average* synonymously. For ungrouped data, the mean is obtained by dividing the sum of all values by the number of values in the data set:

$$\text{Mean} = \frac{\text{Sum of all values}}{\text{Number of values}}$$

The mean calculated for sample data is denoted by $\bar{x}$ (read as "x bar"), and the mean calculated for population data is denoted by μ (Greek letter *mu*). We know from the discussion in Chapter 2 that the number of values in a data set is denoted by n for a sample and by N for a population. In Chapter 1, we learned that a variable is denoted by x, and the sum of all values of x is denoted by Σx. Using these notations, we can write the following formulas for the mean.

Calculating Mean for Ungrouped Data The *mean for ungrouped data* is obtained by dividing the sum of all values by the number of values in the data set. Thus,

$$\text{Mean for population data:} \quad \mu = \frac{\Sigma x}{N}$$

$$\text{Mean for sample data:} \quad \bar{x} = \frac{\Sigma x}{n}$$

where Σx is the sum of all values, N is the population size, n is the sample size, μ is the population mean, and $\bar{x}$ is the sample mean.

■ EXAMPLE 3–1

Calculating the sample mean for ungrouped data.

Table 3.1 lists the total cash donations (rounded to millions of dollars) given by eight U.S. companies during the year 2010 (*Source:* Based on U.S. Internal Revenue Service data analyzed by *The Chronicle of Philanthropy* and *USA TODAY*).

Table 3.1 Cash Donations in 2010 by Eight U.S. Companies

Company	**Cash Donations (millions of dollars)**
Wal-Mart	319
Exxon Mobil	199
Citigroup	110
Home Depot	63
Best Buy	21
Goldman Sachs	315
American Express	26
Nike	63

Find the mean of cash donations made by these eight companies.

Solution The variable in this example is the 2010 cash donations by a company. Let us denote this variable by x. Then, the eight values of x are

$$x_1 = 319, \quad x_2 = 199, \quad x_3 = 110, \quad x_4 = 63,$$
$$x_5 = 21, \quad x_6 = 315, \quad x_7 = 26, \quad \text{and} \quad x_8 = 63$$

where $x_1 = 319$ represents the 2010 cash donations (in millions of dollars) by Wal-Mart, $x_2 = 199$ represents the 2010 cash donations by Exxon Mobil, and so on. The sum of the 2010 cash donations by these eight companies is

$$\begin{aligned}\Sigma x &= x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8\\ &= 319 + 199 + 110 + 63 + 21 + 315 + 26 + 63 = 1116\end{aligned}$$

Note that the given data include only eight companies. Hence, it represents a sample. Because the given data set contains eight companies, $n = 8$. Substituting the values of Σx and n in the sample formula, we obtain the mean of 2010 cash donations of the eight companies as follows:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{1116}{8} = \mathbf{139.5} = \mathbf{\$139.5\ million}$$

Thus, these eight companies donated an average of \$139.5 million in 2010 for charitable purposes. ■

■ EXAMPLE 3–2

Calculating the population mean for ungrouped data.

The following are the ages (in years) of all eight employees of a small company:

53 32 61 27 39 44 49 57

Find the mean age of these employees.

Solution Because the given data set includes *all* eight employees of the company, it represents the population. Hence, $N = 8$. We have

$$\Sigma x = 53 + 32 + 61 + 27 + 39 + 44 + 49 + 57 = 362$$

The population mean is

$$\mu = \frac{\Sigma x}{N} = \frac{362}{8} = \mathbf{45.25\ years}$$

Thus, the mean age of all eight employees of this company is 45.25 years, or 45 years and 3 months. ■

Reconsider Example 3–2. If we take a sample of three employees from this company and calculate the mean age of those three employees, this mean will be denoted by $\bar{x}$. Suppose the three values included in the sample are 32, 39, and 57. Then, the mean age for this sample is

$$\bar{x} = \frac{32 + 39 + 57}{3} = 42.67 \text{ years}$$

If we take a second sample of three employees of this company, the value of $\bar{x}$ will (most likely) be different. Suppose the second sample includes the values 53, 27, and 44. Then, the mean age for this sample is

$$\bar{x} = \frac{53 + 27 + 44}{3} = 41.33 \text{ years}$$

Consequently, we can state that the value of the population mean μ is constant. However, the value of the sample mean $\bar{x}$ varies from sample to sample. The value of $\bar{x}$ for a particular sample depends on what values of the population are included in that sample.

Sometimes a data set may contain a few very small or a few very large values. As mentioned in Chapter 2, such values are called *outliers* or *extreme values.*

A major shortcoming of the mean as a measure of central tendency is that it is very sensitive to outliers. Example 3–3 illustrates this point.

■ EXAMPLE 3–3

Illustrating the effect of an outlier on the mean.

Table 3.2 lists the total number of homes lost to foreclosure in seven states during 2010.

Table 3.2 Number of Homes Foreclosed in 2010

State	Number of Homes Foreclosed
California	173,175
Illinois	49,723
Minnesota	20,352
New Jersey	10,824
Ohio	40,911
Pennsylvania	18,038
Texas	61,848

Note that the number of homes foreclosed in California is very large compared to those in the other six states. Hence, it is an outlier. Show how the inclusion of this outlier affects the value of the mean.

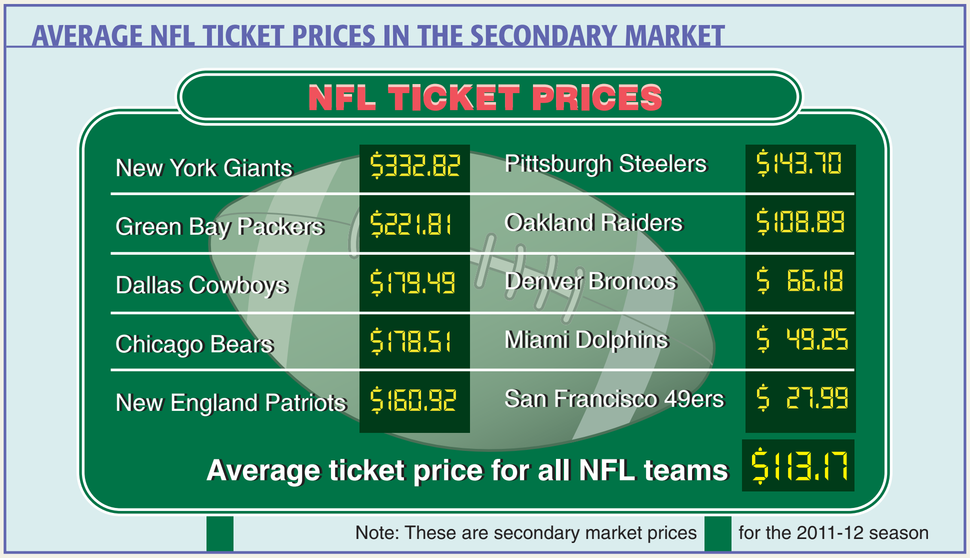
Solution If we do not include the number of homes foreclosed in California (the outlier), the mean of the number of foreclosed homes in six states is

$$\text{Mean without the outlier} = \frac{49{,}723 + 20{,}352 + 10{,}824 + 40{,}911 + 18{,}038 + 61{,}848}{6}$$

$$= \frac{201{,}696}{6} = \mathbf{33{,}616\ homes}$$

CASE STUDY 3–1

AVERAGE NFL TICKET PRICES IN THE SECONDARY MARKET



AVERAGE NFL TICKET PRICES IN THE SECONDARY MARKET

NFL TICKET PRICES

New York Giants	$332.82	Pittsburgh Steelers	$143.70
Green Bay Packers	$221.81	Oakland Raiders	$108.89
Dallas Cowboys	$179.49	Denver Broncos	$ 66.18
Chicago Bears	$178.51	Miami Dolphins	$ 49.25
New England Patriots	$160.92	San Francisco 49ers	$ 27.99
		Average ticket price for all NFL teams	$113.17

Note: These are secondary market prices for the 2011-12 season

Data source: seatgeek.com.

The accompanying chart, based on data from seatgeek.com, shows the 2011–2012 average secondary market ticket price for all NFL teams as well as for a selected number of NFL teams. (Note that the secondary market tickets are the tickets that are resold via ticket Web sites such as Seatgeek.com. These tickets are not purchased directly from NFL franchises.) According to the data on seatgeek.com, the New York Giants had the highest secondary market average ticket price at $332.82 and the San Francisco 49ers had the lowest secondary market average ticket price at $27.99 for the 2011–2012 season. As we can see from the chart, there is a huge variation in the secondary market average ticket prices for these 10 teams, and this is true for all NFL teams. The secondary market average ticket price for all NFL teams was $113.17 during the 2011–2012 season.

Source: http://seatgeek.com/blog/nfl/average-ticket-prices-nfl.

Now, to see the impact of the outlier on the value of the mean, we include the number of homes foreclosed in California and find the mean number of homes foreclosed in the seven states. This mean is

$$\text{Mean with the outlier} = \frac{173{,}175 + 49{,}723 + 20{,}352 + 10{,}824 + 40{,}911 + 18{,}038 + 61{,}848}{7}$$

$$= \frac{374{,}871}{7} = \mathbf{53{,}553}$$

Thus, including the number of homes foreclosed in California increases the mean by about 60% from 33,616 to 53,553. ■

The preceding example should encourage us to be cautious. We should remember that the mean is not always the best measure of central tendency because it is heavily influenced by outliers. Sometimes other measures of central tendency give a more accurate impression of a data set. For example, when a data set has outliers, instead of using the mean, we can use either the trimmed mean (defined in Exercise 3.33) or the median (to be discussed next) as a measure of central tendency.

3.1.2 Median

Another important measure of central tendency is the **median**. It is defined as follows.

Definition

Median The *median* is the value of the middle term in a data set that has been ranked in increasing order.

CASE STUDY 3–2

AVERAGE IS OVER

By Thomas L. Friedman. The following article was originally published in The New York Times.

In an essay, entitled "Making It in America," in the latest issue of the Atlantic, the author Adam Davidson relates a joke from cotton country about just how much a modern textile mill has been automated: The average mill has only two employees today: "a man and a dog. The man is there to feed the dog, and the dog is there to keep the man away from the machines."

Davidson's article is one of a number of pieces that have recently appeared making the point that the reason we have such stubbornly high unemployment and sagging middle-class incomes today is largely because of the big drop in demand because of the Great Recession, but it is also because of the quantum advances in both globalization and the information technology revolution, which are more rapidly than ever replacing labor with machines or foreign workers.

In the past, workers with average skills, doing an average job, could earn an average lifestyle. But, today, average is officially over. Being average just won't earn you what it used to. It can't when so many more employers have so much more access to so much more above average cheap foreign labor, cheap robotics, cheap software, cheap automation and cheap genius. Therefore, everyone needs to find their extra—their unique value contribution that makes them stand out in whatever is their field of employment. *Average is over*.

Yes, new technology has been eating jobs forever, and always will. As they say, if horses could have voted, there never would have been cars. But there's been an acceleration. As Davidson notes, "In the 10 years ending in 2009, [U.S.] factories shed workers so fast that they erased almost all the gains of the previous 70 years; roughly one out of every three manufacturing jobs—about 6 million in total—disappeared."

And you ain't seen nothin' yet. Last April, Annie Lowrey of Slate wrote about a start-up called "E la Carte" that is out to shrink the need for waiters and waitresses: The company "has produced a kind of souped-up iPad that lets you order and pay right at your table. The brainchild of a bunch of M.I.T. engineers, the nifty invention, known as the Presto, might be found at a restaurant near you soon . . . You select what you want to eat and add items to a cart. Depending on the restaurant's preferences, the console could show you nutritional information, ingredients lists and photographs. You can make special requests, like 'dressing on the side' or 'quintuple bacon.' When you're done, the order zings over to the kitchen, and the Presto tells you how long it will take for your items to come out. . . . Bored with your companions? Play games on the machine. When you're through with your meal, you pay on the console, splitting the bill item by item if you wish and paying however you want. And you can have your receipt emailed to you . . . Each console goes for $100 per month. If a restaurant serves meals eight hours a day, seven days a week, it works out to 42 cents per hour per table—making the Presto cheaper than even the very cheapest waiter.

What the iPad won't do in an above average way a Chinese worker will. Consider this paragraph from Sunday's terrific article in The Times by Charles Duhigg and Keith Bradsher about why Apple does so much of its manufacturing in China: "Apple had redesigned the iPhone's screen at the last minute, forcing an assembly-line overhaul. New screens began arriving at the [Chinese] plant near midnight. A foreman immediately roused 8,000 workers inside the company's dormitories, according to the executive. Each employee was given a biscuit and a cup of tea, guided to a workstation and within half an hour started a 12-hour shift fitting glass screens into beveled frames. Within 96 hours, the plant was producing over 10,000 iPhones a day. 'The speed and flexibility is breathtaking,' the executive said. 'There's no American plant that can match that.' "

And automation is not just coming to manufacturing, explains Curtis Carlson, the chief executive of SRI International, a Silicon Valley idea lab that invented the Apple iPhone program known as Siri, the digital personal assistant. "Siri is the beginning of a huge transformation in how we interact with banks, insurance companies, retail stores, health care providers, information retrieval services and product services."

There will always be change—new jobs, new products, new services. But the one thing we know for sure is that with each advance in globalization and the I.T. revolution, the best jobs will require workers to have more and better education to make themselves above average. Here are the latest unemployment rates from the Bureau of Labor Statistics for Americans over 25 years old: those with less than a high school degree, 13.8 percent; those with a high school degree and no college, 8.7 percent; those with some college or associate degree, 7.7 percent; and those with bachelor's degree or higher, 4.1 percent.

In a world where average is officially over, there are many things we need to do to buttress employment, but nothing would be more important than passing some kind of G.I. Bill for the 21st century that ensures that every American has access to post-high school education.

Source: Thomas L. Friedman, *The New York Times*, January 25, 2012. Copyright © 2012, *The New York* Times. All rights reserved. Used with permission and protected by the Copyright Laws of the United States. The printing, copying, redistribution, or retransmission of this Content without express written permission is prohibited.

As is obvious from the definition of the median, it divides a ranked data set into two equal parts. The calculation of the median consists of the following two steps:

1. Rank the data set in increasing order.
2. Find the middle term. The value of this term is the median.[1]

Note that if the number of observations in a data set is *odd*, then the median is given by the value of the middle term in the ranked data. However, if the number of observations is *even*, then the median is given by the average of the values of the two middle terms.

■ EXAMPLE 3–4

Calculating the median for ungrouped data: odd number of data values.

Refer to the data on the number of homes foreclosed in seven states given in Table 3.2 of Example 3–3. Those values are listed below.

173,175 49,723 20,352 10,824 40,911 18,038 61,848

Find the median for these data.

Solution First, we rank the given data in increasing order as follows:

10,824 18,038 20,352 40,911 49,723 61,848 173,175

Since there are seven homes in this data set and the middle term is the fourth term, the median is given by the value of the fourth term in the ranked data as shown below.

10,824 18,038 20,352 40,911 49,723 61,848 173,175
↑
Median

Thus, the median number of homes foreclosed in these seven states was 40,911 in 2010. ■

■ EXAMPLE 3–5

Calculating the median for ungrouped data: even number of data values.

Table 3.3 gives the total compensations (in millions of dollars) for the year 2010 of the 12 highest-paid CEOs of U.S. companies.

Table 3.3 Total Compensations of 12 Highest-Paid CEOs for the Year 2010

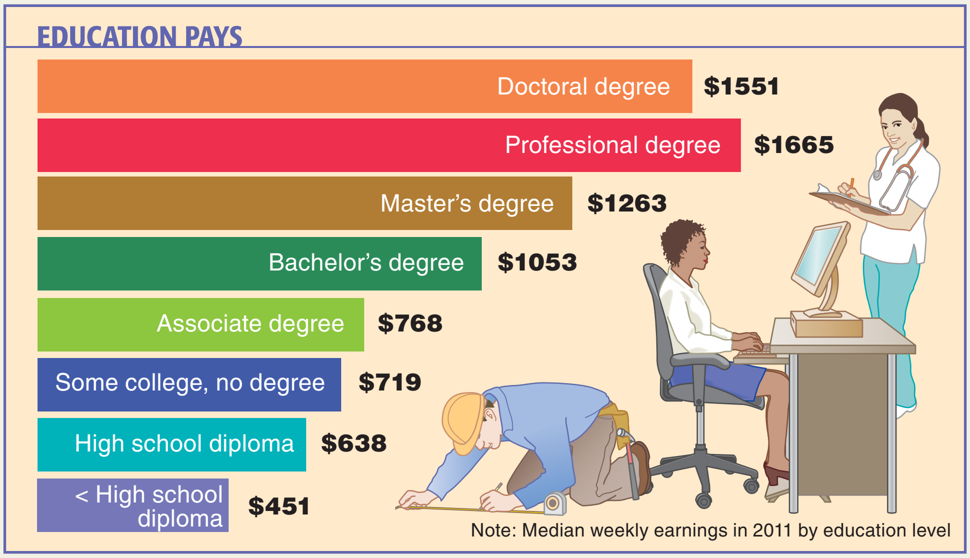
CEO and Company	2010 Total Compensation (millions of dollars)
Michael D. White (DirecTV)	32.9
David N. Farr (Emerson Electric)	22.9
Brian L. Roberts (Comcast)	28.2
Philippe P. Dauman (Viacom)	84.5
William C. Weldon (Johnson & Johnson)	21.6
Robert A. Iger (Walt Disney)	28.0
Ray R. Iran (Occidental Petroleum)	76.1
Samuel J. Palmisano (IBM)	25.2
John F. Lundgren (Stanley Black & Decker)	32.6
Lawrence J. Ellison (Oracle)	70.1
Alan Mulally (Ford Motor)	26.5
Howard Schultz (Starbucks)	21.7

Find the median for these data.

[1] The value of the middle term in a data set ranked in *decreasing* order will also give the value of the median.

CASE STUDY 3–3

EDUCATION PAYS



Data source: U.S. Bureau of Labor Statistics.

Data source: http://www.bls.gov/emp/ep_chart_001.htm/.

The accompanying chart shows the 2011 median weekly salaries by education level for persons of age 25 years and older who held full-time jobs. These salaries are based on the Current Population Survey conducted by the Bureau of Labor Statistics. Although this survey is called the Current Population Survey, it is actually based on a sample. Usually the samples taken by the Bureau of Labor Statistics for these surveys are very large. As shown in the chart, the highest median weekly earning (of $1665) was for workers with a professional degree and the lowest (of $451) was for workers with less than a high school diploma.

Solution First we rank the given total compensations of the 12 CEOs as follows:

21.6 21.7 22.9 25.2 26.5 28.0 28.2 32.6 32.9 70.1 76.1 84.5

There are 12 values in this data set. Because there is an even number of values in the data set, the median will be given by the average of the two middle values. The two middle values are the sixth and seventh in the arranged data, and these two values are 28.0 and 28.2. The median, which is given by the average of these two values, is calculated as follows:

$$21.6 \quad 21.7 \quad 22.9 \quad 25.2 \quad 26.5 \quad \boxed{\underset{\text{Median } = \text{ } 28.1}{\underset{\uparrow}{28.0 \quad 28.2}}} \quad 32.6 \quad 32.9 \quad 70.1 \quad 76.1 \quad 84.5$$

$$\text{Median} = \frac{28.0 + 28.2}{2} = \frac{56.2}{2} = 28.1 = \textbf{\$28.1 million}$$

Thus, the median for the 2010 compensations of these 12 CEOs is $28.1 million. ■

The median gives the center of a histogram, with half of the data values to the left of the median and half to the right of the median. The advantage of using the median as a measure of central tendency is that it is not influenced by outliers. Consequently, the median is preferred over the mean as a measure of central tendency for data sets that contain outliers.

3.1.3 Mode

Mode is a French word that means *fashion*—an item that is most popular or common. In statistics, the mode represents the most common value in a data set.

Definition

Mode The *mode* is the value that occurs with the highest frequency in a data set.

■ EXAMPLE 3–6

Calculating the mode for ungrouped data.

The following data give the speeds (in miles per hour) of eight cars that were stopped on I-95 for speeding violations.

77	82	74	81	79	84	74	78

Find the mode.

Solution In this data set, 74 occurs twice, and each of the remaining values occurs only once. Because 74 occurs with the highest frequency, it is the mode. Therefore,

$$\text{Mode} = \textbf{74 miles per hour}$$ ■

A major shortcoming of the mode is that a data set may have none or may have more than one mode, whereas it will have only one mean and only one median. For instance, a data set with each value occurring only once has no mode. A data set with only one value occurring with the highest frequency has only one mode. The data set in this case is called **unimodal**. A data set with two values that occur with the same (highest) frequency has two modes. The distribution, in this case, is said to be **bimodal**. If more than two values in a data set occur with the same (highest) frequency, then the data set contains more than two modes and it is said to be **multimodal**.

■ EXAMPLE 3–7

Data set with no mode.

Last year's incomes of five randomly selected families were $76,150, $95,750, $124,985, $87,490, and $53,740. Find the mode.

Solution Because each value in this data set occurs only once, this data set contains **no mode**. ■

■ EXAMPLE 3–8

Data set with two modes.

A small company has 12 employees. Their commuting times (rounded to the nearest minute) from home to work are 23, 36, 12, 23, 47, 32, 8, 12, 26, 31, 18, and 28, respectively. Find the mode for these data.

Solution In the given data on the commuting times of these 12 employees, each of the values 12 and 23 occurs twice, and each of the remaining values occurs only once. Therefore, this data set has two modes: 12 and 23 minutes. ■

■ EXAMPLE 3–9

Data set with three modes.

The ages of 10 randomly selected students from a class are 21, 19, 27, 22, 29, 19, 25, 21, 22, and 30 years, respectively. Find the mode.

Solution This data set has three modes: **19**, **21**, and **22**. Each of these three values occurs with a (highest) frequency of 2. ■

One advantage of the mode is that it can be calculated for both kinds of data—quantitative and qualitative—whereas the mean and median can be calculated for only quantitative data.

■ EXAMPLE 3–10

Finding the mode for qualitative data.

The status of five students who are members of the student senate at a college are senior, sophomore, senior, junior, and senior, respectively. Find the mode.

Solution Because **senior** occurs more frequently than the other categories, it is the mode for this data set. We cannot calculate the mean and median for this data set. ■

To sum up, we cannot say for sure which of the three measures of central tendency is a better measure overall. Each of them may be better under different situations. Probably the mean is the most-used measure of central tendency, followed by the median. The mean has the advantage that its calculation includes each value of the data set. The median is a better measure when a data set includes outliers. The mode is simple to locate, but it is not of much use in practical applications.

3.1.4 Relationships Among the Mean, Median, and Mode

As discussed in Chapter 2, two of the many shapes that a histogram or a frequency distribution curve can assume are symmetric and skewed. This section describes the relationships among the mean, median, and mode for three such histograms and frequency distribution curves. Knowing the values of the mean, median, and mode can give us some idea about the shape of a frequency distribution curve.

1. For a symmetric histogram and frequency distribution curve with one peak (see Figure 3.2), the values of the mean, median, and mode are identical, and they lie at the center of the distribution.

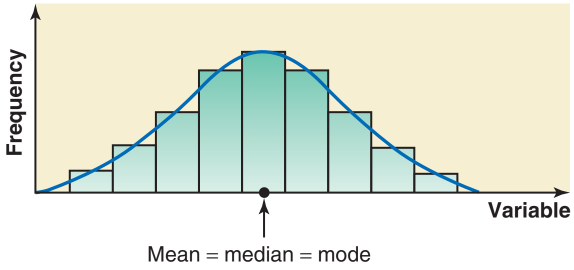


Figure 3.2 Mean, median, and mode for a symmetric histogram and frequency distribution curve.

2. For a histogram and a frequency distribution curve skewed to the right (see Figure 3.3), the value of the mean is the largest, that of the mode is the smallest, and the value of the median lies between these two. (Notice that the mode always occurs at the peak point.) The value of the mean is the largest in this case because it is sensitive to outliers that occur in the right tail. These outliers pull the mean to the right.

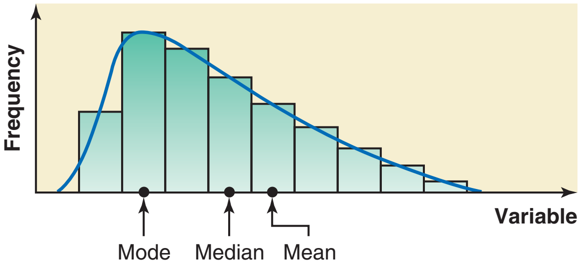


Figure 3.3 Mean, median, and mode for a histogram and frequency distribution curve skewed to the right.

3. If a histogram and a frequency distribution curve are skewed to the left (see Figure 3.4), the value of the mean is the smallest and that of the mode is the largest, with the value of the median lying between these two. In this case, the outliers in the left tail pull the mean to the left.

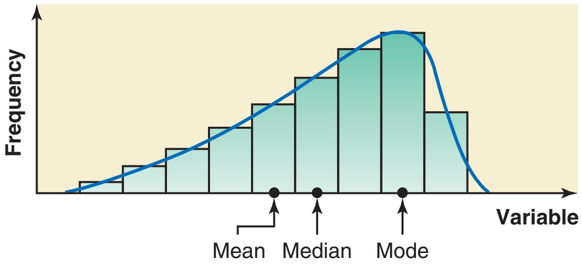


Figure 3.4 Mean, median, and mode for a histogram and frequency distribution curve skewed to the left.

EXERCISES

CONCEPTS AND PROCEDURES

3.1 Explain how the value of the median is determined for a data set that contains an odd number of observations and for a data set that contains an even number of observations.

3.2 Briefly explain the meaning of an outlier. Is the mean or the median a better measure of central tendency for a data set that contains outliers? Illustrate with the help of an example.

3.3 Using an example, show how outliers can affect the value of the mean.

3.4 Which of the three measures of central tendency (the mean, the median, and the mode) can be calculated for quantitative data only, and which can be calculated for both quantitative and qualitative data? Illustrate with examples.

3.5 Which of the three measures of central tendency (the mean, the median, and the mode) can assume more than one value for a data set? Give an example of a data set for which this summary measure assumes more than one value.

3.6 Is it possible for a (quantitative) data set to have no mean, no median, or no mode? Give an example of a data set for which this summary measure does not exist.

3.7 Explain the relationships among the mean, median, and mode for symmetric and skewed histograms. Illustrate these relationships with graphs.

3.8 Prices of cars have a distribution that is skewed to the right with outliers in the right tail. Which of the measures of central tendency is the best to summarize this data set? Explain.

3.9 The following data set belongs to a population:

$$4 \qquad -7 \qquad 1 \qquad 0 \qquad -9 \qquad 16 \qquad 9 \qquad 8$$

Calculate the mean, median, and mode.

3.10 The following data set belongs to a sample:

$$12 \qquad 4 \qquad -10 \qquad 8 \qquad 8 \qquad -13$$

Calculate the mean, median, and mode.

APPLICATIONS

3.11 The following table gives the standard deductions and personal exemptions for persons filing with "single" status on their 2011 state income taxes in a random sample of 9 states. Calculate the mean and median for the data on standard deductions for these states.

State	Standard Deduction (in dollars)	Personal Exemption (in dollars)
Delaware	3250	110
Hawaii	2000	1040
Kentucky	2190	20
Minnesota	5450	3500
North Dakota	5700	3650
Oregon	1945	176
Rhode Island	5700	3650
Vermont	5700	3650
Virginia	3000	930

Source: www.taxfoundation.org.

3.12 Refer to the data table in Exercise 3.11. Calculate the mean and median for the data on personal exemptions for these states.

3.13 The following data give the 2010 gross domestic product (in billions of dollars) for all 50 states. The data are entered in alphabetical order by state (*Source*: Bureau of Economic Analysis).

173	49	254	103	1901	258	237	62	748	403
67	55	652	276	143	127	163	219	52	295
379	384	270	97	244	36	90	126	60	487
80	1160	425	35	478	148	174	570	49	164
40	255	1207	115	26	424	340	65	248	39

a. Calculate the mean and median for these data. Are these values of the mean and the median sample statistics or population parameters? Explain.
b. Do these data have a mode? Explain.

3.14 The following data give the 2010 revenues (in millions of dollars) of the six Maryland-based companies listed in the 2010 *Fortune 500* (www.money.cnn.com/magazines/fortune/fortune500/2010/states/MD.html). The data refer to the following companies, respectively: Lockheed Martin, Constellation Energy, Coventry Health Care, Marriott International, Black & Decker, and Host Hotels & Resorts.

45,189.0 15,598.8 13,993.3 10,908.0 4775.1 4216.0

Find the mean and median for these data. Do these data have a mode? Assume that these six companies constitute the population of companies from Maryland in the 2010 *Fortune 500*.

3.15 The following table lists the 2009 total profits (rounded to millions of dollars) of the seven *Fortune 500* companies in the Computers, Office Equipment category (*Source:* www.money.cnn.com/magazines/fortune/fortune500/2010/industries/8/index.html).

Company	**2009 Profit (millions of dollars)**
Hewlett-Packard	7660
Dell	1433
Apple	5704
Xerox	485
Sun Microsystems	−2234
Pitney Bowes	423
NCR	−33

Find the mean and median for these data. (Note: The negative values for Sun Microsystems and NCR imply that both companies lost money in 2009.) Assume that these seven companies constitute the population of such companies in the 2009 *Fortune 500*.

3.16 The following data give the numbers of car thefts that occurred in a city during the past 14 days.

8 3 7 11 4 3 5 4 2 6 4 15 8 3

Find the mean, median, and mode.

3.17 The following data give the amount of money (in dollars) that each of six Canadian social service charities spent to raise $100 in donations during 2010 (www.moneysense.ca). The values, listed in that order, are for the Calgary Inter-Faith Food Bank Society, Covenant House Toronto, The Salvation Army Territorial Headquarters for Canada and Bermuda, Second Harvest Food Support Committee, Teen Challenge, and Toronto Windfall Clothing Support Service.

.20 29.30 11.30 5.30 9.90 .50

Compute the mean and median. Do these data have a mode? Why or why not?

3.18 The following table gives the number of major penalties for each of the 15 teams in the Eastern Conference of the National Hockey League during the 2010–11 season (www.nhl.com). A major penalty is subject to 5 minutes in the penalty box for a player.

Team	Number of Major Penalties
Pittsburgh	74
Boston	73
New York Islanders	71
New York Rangers	62
Columbus	59
Toronto	53
Ottawa	51
Philadelphia	49
Washington	46
New Jersey	39
Montreal	35
Atlanta	34
Buffalo	30
Florida	26
Tampa Bay	23

Compute the mean and median for the data on major penalties. Do these data have a mode? Why or why not?

3.19 Due to antiquated equipment and frequent windstorms, the town of Oak City often suffers power outages. The following data give the numbers of power outages for each of the past 12 months.

4 5 7 3 2 0 2 3 2 1 2 4

Compute the mean, median, and mode for these data.

3.20 Standard milk chocolate M&Ms™ come in six colors. The Fun Size bags typically contain between 16 and 20 candies, so it is common for a Fun Size bag to have some of the six colors missing. Each of the 14 students in a summer statistics class was given a Fun Size bag and asked to count the number of colors present in the bag. The following data are the number of colors found in these 14 bags:

5 6 4 4 6 3 2 4 5 4 3 6 3 6

Find the mean, median, and mode for these data. Are the values of these summary measures population parameters or sample statistics? Explain why.

3.21 Nixon Corporation manufactures computer monitors. The following data are the numbers of computer monitors produced at the company for a sample of 10 days.

24 31 27 25 35 33 26 40 25 28

Calculate the mean, median, and mode for these data.

3.22 Grand Jury indictment data for Gloucester County, New Jersey, are published every week in the *Gloucester County Times* newspaper (www.nj.com/gloucester). The following data are the number of indictments for a sample of 11 weeks selected from July 2010 through June 2011:

35 13 17 21 21 29 20 26 24 13 23

Find the mean, median, and mode for these data.

3.23 The following data represent the numbers of tornadoes that touched down during 1950 to 1994 in the 12 states that had the most tornadoes during this period. The data for these states are given in the following order: CO, FL, IA, IL, KS, LA, MO, MS, NE, OK, SD, TX.

1113 2009 1374 1137 2110 1086 1166 1039 1673 2300 1139 5490

a. Calculate the mean and median for these data.
b. Identify the outlier in this data set. Drop the outlier and recalculate the mean and median. Which of these two summary measures changes by a larger amount when you drop the outlier?
c. Which is the better summary measure for these data, the mean or the median? Explain.

3.24 The following data set lists the number of women from each of 12 countries who were on the Rolex Women's World Golf Rankings Top 50 list as of July 18, 2011. The data, listed in that order, are for the following countries: Australia, Chinese Taipei, England, Germany, Japan, Netherlands, Norway, Scotland, South Korea, Spain, Sweden, and the United States.

3	1	1	1	10	1	1	1	18	2	3	8

a. Calculate the mean and median for these data.
b. Identify the outlier in this data set. Drop the outlier and recalculate the mean and median. Which of these two summary measures changes by a larger amount when you drop the outlier?
c. Which is the better summary measure for these data, the mean or the median? Explain.

***3.25** One property of the mean is that if we know the means and sample sizes of two (or more) data sets, we can calculate the **combined mean** of both (or all) data sets. The combined mean for two data sets is calculated by using the formula

$$\text{Combined mean} = \bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

where n_1 and n_2 are the sample sizes of the two data sets and $\bar{x}_1$ and $\bar{x}_2$ are the means of the two data sets, respectively. Suppose a sample of 10 statistics books gave a mean price of \$140 and a sample of 8 mathematics books gave a mean price of \$160. Find the combined mean. (*Hint:* For this example: $n_1 = 10, n_2 = 8, \bar{x}_1 = \$140, \bar{x}_2 = \$160$.)

***3.26** Twenty business majors and 15 economics majors go bowling. Each student bowls one game. The scorekeeper announces that the mean score for the 15 economics majors is 138 and the mean score for the entire group of 35 students is 150. Find the mean score for the 20 business majors.

***3.27** For any data, the sum of all values is equal to the product of the sample size and mean; that is, $\Sigma x = n\bar{x}$. Suppose the average amount of money spent on shopping by 10 persons during a given week is \$105.50. Find the total amount of money spent on shopping by these 10 persons.

***3.28** The mean 2011 income for five families was \$99,520. What was the total 2011 income of these five families?

***3.29** The mean age of six persons is 49 years. The ages of five of these six persons are 55, 39, 44, 51, and 45 years, respectively. Find the age of the sixth person.

***3.30** Seven airline passengers in economy class on the same flight paid an average of \$418 per ticket. Because the tickets were purchased at different times and from different sources, the prices varied. The first five passengers paid \$420, \$210, \$415, \$695, and \$496. The sixth and seventh tickets were purchased by a couple who paid identical fares. What price did each of them pay?

***3.31** Consider the following two data sets.

Data Set I:	12	25	37	8	41
Data Set II:	19	32	44	15	48

Notice that each value of the second data set is obtained by adding 7 to the corresponding value of the first data set. Calculate the mean for each of these two data sets. Comment on the relationship between the two means.

***3.32** Consider the following two data sets.

Data Set I:	4	8	15	9	11
Data Set II:	12	24	45	27	33

Notice that each value of the second data set is obtained by multiplying the corresponding value of the first data set by 3. Calculate the mean for each of these two data sets. Comment on the relationship between the two means.

***3.33** The **trimmed mean** is calculated by dropping a certain percentage of values from each end of a ranked data set. The trimmed mean is especially useful as a measure of central tendency when a data set contains a few outliers. Suppose the following data give the ages (in years) of 10 employees of a company:

47	53	38	26	39	49	19	67	31	23

To calculate the 10% trimmed mean, first rank these data values in increasing order; then drop 10% of the smallest values and 10% of the largest values. The mean of the remaining 80% of the values will give the 10% trimmed mean. Note that this data set contains 10 values, and 10% of 10 is 1. Thus, if we drop the smallest value and the largest value from this data set, the mean of the remaining 8 values will be called the 10% trimmed mean. Calculate the 10% trimmed mean for this data set.

***3.34** The following data give the prices (in thousands of dollars) of 20 houses sold recently in a city.

184	324	365	309	245	387	369	438	195	180
323	578	510	679	307	271	318	795	259	590

Find the 20% trimmed mean for this data set.

3.35** In some applications, certain values in a data set may be considered more important than others. For example, to determine students' grades in a course, an instructor may assign a weight to the final exam that is twice as much as that to each of the other exams. In such cases, it is more appropriate to use the **weighted mean**. In general, for a sequence of n data values $x_1, x_2, \dots, x_n$ that are assigned weights $w_1, w_2, \dots, w_n$, respectively, the ***weighted mean is found by the formula

$$\text{Weighted mean} = \frac{\Sigma xw}{\Sigma w}$$

where Σxw is obtained by multiplying each data value by its weight and then adding the products. Suppose an instructor gives two exams and a final, assigning the final exam a weight twice that of each of the other exams. Find the weighted mean for a student who scores 73 and 67 on the first two exams and 85 on the final. (*Hint:* Here, $x_1 = 73$, $x_2 = 67$, $x_3 = 85$, $w_1 = w_2 = 1$, and $w_3 = 2$.)

***3.36** When studying phenomena such as inflation or population changes that involve periodic increases or decreases, the **geometric mean** is used to find the average change over the entire period under study. To calculate the geometric mean of a sequence of n values $x_1, x_2, \dots, x_n$, we multiply them together and then find the nth root of this product. Thus

$$\text{Geometric mean} = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \ \dots \ \cdot x_n}$$

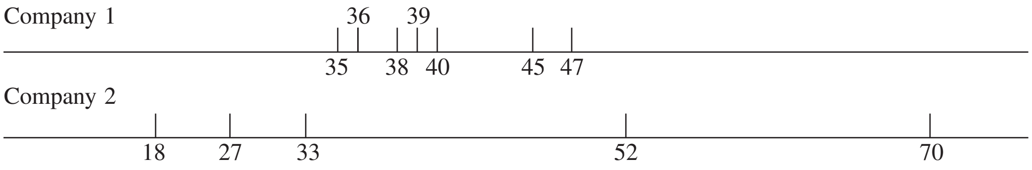
Suppose that the inflation rates for the last five years are 4%, 3%, 5%, 6%, and 8%, respectively. Thus at the end of the first year, the price index will be 1.04 times the price index at the beginning of the year, and so on. Find the mean rate of inflation over the 5-year period by finding the geometric mean of the data set 1.04, 1.03, 1.05, 1.06, and 1.08. (*Hint:* Here, $n = 5$, $x_1 = 1.04$, $x_2 = 1.03$, and so on. Use the $x^{1/n}$ key on your calculator to find the fifth root. Note that the mean inflation rate will be obtained by subtracting 1 from the geometric mean.)

3.2 Measures of Dispersion for Ungrouped Data

The measures of central tendency, such as the mean, median, and mode, do not reveal the whole picture of the distribution of a data set. Two data sets with the same mean may have completely different spreads. The variation among the values of observations for one data set may be much larger or smaller than for the other data set. (Note that the words *dispersion*, *spread*, and *variation* have similar meanings.) Consider the following two data sets on the ages (in years) of all workers working for each of two small companies.

Company 1:	47	38	35	40	36	45	39
Company 2:		70	33	18	52	27	

The mean age of workers in both these companies is the same, 40 years. If we do not know the ages of individual workers at these two companies and are told only that the mean age of the workers at both companies is the same, we may deduce that the workers at these two companies have a similar age distribution. As we can observe, however, the variation in the workers' ages for each of these two companies is very different. As illustrated in the diagram, the ages of the workers at the second company have a much larger variation than the ages of the workers at the first company.



Thus, the mean, median, or mode by itself is usually not a sufficient measure to reveal the shape of the distribution of a data set. We also need a measure that can provide some information about the variation among data values. The measures that help us learn about the spread of a data set are called the **measures of dispersion**. The measures of central tendency and dispersion taken together give a better picture of a data set than the measures of central tendency alone. This section discusses three measures of dispersion: range, variance, and standard deviation. Another measure of spread, called the coefficient of variation (CV), is explained in Exercise 3.57.

3.2.1 Range

The **range** is the simplest measure of dispersion to calculate. It is obtained by taking the difference between the largest and the smallest values in a data set.

Finding the Range for Ungrouped Data

$$\text{Range} = \text{Largest value} - \text{Smallest value}$$

■ EXAMPLE 3–11

Calculating the range for ungrouped data.

Table 3.4 gives the total areas in square miles of the four western South-Central states of the United States.

Table 3.4

State	Total Area (square miles)
Arkansas	53,182
Louisiana	49,651
Oklahoma	69,903
Texas	267,277

Find the range for this data set.

Solution The maximum total area for a state in this data set is 267,277 square miles, and the smallest area is 49,651 square miles. Therefore,

$$\begin{aligned}\text{Range} &= \text{Largest value} - \text{Smallest value}\\ &= 267{,}277 - 49{,}651 = \mathbf{217{,}626\ square\ miles}\end{aligned}$$

Thus, the total areas of these four states are spread over a range of 217,626 square miles. ■

The range, like the mean, has the disadvantage of being influenced by outliers. In Example 3–11, if the state of Texas with a total area of 267,277 square miles is dropped, the range decreases from 217,626 square miles to 20,252 square miles. Consequently, the range is not a good measure of dispersion to use for a data set that contains outliers.

Another disadvantage of using the range as a measure of dispersion is that its calculation is based on two values only: the largest and the smallest. All other values in a data set are ignored when calculating the range. Thus, the range is not a very satisfactory measure of dispersion.

3.2.2 Variance and Standard Deviation

The **standard deviation** is the most-used measure of dispersion. The value of the standard deviation tells how closely the values of a data set are clustered around the mean. In general, a lower value of the standard deviation for a data set indicates that the values of that data set are spread over a relatively smaller range around the mean. In contrast, a larger value of the standard deviation for a data set indicates that the values of that data set are spread over a relatively larger range around the mean.

The *standard deviation is obtained by taking the positive square root of the* **variance**. The variance calculated for population data is denoted by $\boldsymbol{\sigma^2}$ (read as *sigma squared*),[2] and the variance calculated for sample data is denoted by $\boldsymbol{s^2}$. Consequently, the standard deviation

[2] Note that Σ is uppercase sigma and σ is lowercase sigma of the Greek alphabet.

calculated for population data is denoted by $\boldsymbol{\sigma}$, and the standard deviation calculated for sample data is denoted by $\boldsymbol{s}$. Following are what we will call the *basic formulas* that are used to calculate the variance and standard deviation.[3]

$$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N} \quad \text{and} \quad s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$$

$$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}} \quad \text{and} \quad s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$

where σ^2 is the population variance, s^2 is the sample variance, σ is the population standard deviation, and s is the sample standard deviation.

The quantity $x - \mu$ or $x - \bar{x}$ in the above formulas is called the *deviation* of the x value from the mean. The sum of the deviations of the x values from the mean is always zero; that is, $\Sigma(x - \mu) = 0$ and $\Sigma(x - \bar{x}) = 0$.

For example, suppose the midterm scores of a sample of four students are 82, 95, 67, and 92, respectively. Then, the mean score for these four students is

$$\bar{x} = \frac{82 + 95 + 67 + 92}{4} = 84$$

The deviations of the four scores from the mean are calculated in Table 3.5. As we can observe from the table, the sum of the deviations of the x values from the mean is zero; that is, $\Sigma(x - \bar{x}) = 0$. For this reason we square the deviations to calculate the variance and standard deviation.

Table 3.5

x	$x - \bar{x}$
82	$82 - 84 = -2$
95	$95 - 84 = +11$
67	$67 - 84 = -17$
92	$92 - 84 = +8$
	$\Sigma(x - \bar{x}) = 0$

From the computational point of view, it is easier and more efficient to use *short-cut formulas* to calculate the variance and standard deviation. By using the short-cut formulas, we reduce the computation time and round-off errors. Use of the basic formulas for ungrouped data is illustrated in Section A3.1.1 of Appendix 3.1 of this chapter. The short-cut formulas for calculating the variance and standard deviation are as follows.

Short-Cut Formulas for the Variance and Standard Deviation for Ungrouped Data

$$\sigma^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{N}}{N} \quad \text{and} \quad s^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}$$

where σ^2 is the population variance and s^2 is the sample variance.

The standard deviation is obtained by taking the positive square root of the variance.

Population standard deviation: $\sigma = \sqrt{\dfrac{\Sigma x^2 - \frac{(\Sigma x)^2}{N}}{N}}$

Sample standard deviation: $s = \sqrt{\dfrac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}}$

[3] From the formula for σ^2, it can be stated that the population variance is the mean of the squared deviations of x values from the mean. However, this is not true for the variance calculated for a sample data set.

Note that the denominator in the formula for the population variance is N, but that in the formula for the sample variance it is $n - 1$.[4]

EXAMPLE 3–12

Calculating the sample variance and standard deviation for ungrouped data.

Until about 2009, airline passengers were not charged for checked baggage. Around 2009, however, many U.S. airlines started charging a fee for bags. According to the Bureau of Transportation Statistics, U.S. airlines collected more than $3 billion in baggage fee revenue in 2010. The following table lists the baggage fee revenues of six U.S. airlines for the year 2010. (Note that Delta's revenue reflects a merger with Northwest. Also note that since then United and Continental have merged; and American filed for bankruptcy and may merge with another airline.)

© Hazlan Abdul Hakim/Stockphoto

Airline	Baggage Fee Revenue (millions of dollars)
United	313
Continental	342
American	581
Delta	952
US Airways	514
AirTran	152

Find the variance and standard deviation for these data.

Solution Let x denote the 2010 baggage fee revenue (in millions of dollars) of an airline. The values of Σx and Σx^2 are calculated in Table 3.6.

Table 3.6

x	x^2
313	97,969
342	116,964
581	337,561
952	906,304
514	264,196
152	23,104
$\Sigma x = 2854$	$\Sigma x^2 = 1{,}746{,}098$

Calculation of the variance and standard deviation involves the following four steps.

Step 1. *Calculate Σx.*

The sum of the values in the first column of Table 3.6 gives the value of Σx, which is 2854.

Step 2. *Find Σx^2.*

The value of Σx^2 is obtained by squaring each value of x and then adding the squared values. The results of this step are shown in the second column of Table 3.6. Notice that $\Sigma x^2 = 1{,}746{,}098$.

[4] The reason that the denominator in the sample formula is $n - 1$ and not n follows: The sample variance underestimates the population variance when the denominator in the sample formula for variance is n. However, the sample variance does not underestimate the population variance if the denominator in the sample formula for variance is $n - 1$. In Chapter 8 we will learn that $n - 1$ is called the degrees of freedom.

Step 3. *Determine the variance.*

Substitute all the values in the variance formula and simplify. Because the given data are for the baggage fee revenues of only six airlines, we use the formula for the sample variance:

$$s^2 = \frac{\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}}{n - 1} = \frac{1{,}746{,}098 - \dfrac{(2854)^2}{6}}{6 - 1} = \frac{1{,}746{,}098 - 1{,}357{,}552.667}{5} = \mathbf{77{,}709.06666}$$

Step 4. *Obtain the standard deviation.*

The standard deviation is obtained by taking the (positive) square root of the variance:

$$s = \sqrt{\frac{\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}}{n - 1}} = \sqrt{77{,}709.06666} = 278.7634601 = \mathbf{\$278.76\ million}$$

Thus, the standard deviation of the 2010 baggage fee revenues of these six airlines is $278.76 million. ■

◀ *Two Observations*

1. **The values of the variance and the standard deviation are never negative**. That is, the numerator in the formula for the variance should never produce a negative value. Usually the values of the variance and standard deviation are positive, but if a data set has no variation, then the variance and standard deviation are both zero. For example, if four persons in a group are the same age—say, 35 years—then the four values in the data set are

   35 35 35 35

   If we calculate the variance and standard deviation for these data, their values are zero. This is because there is no variation in the values of this data set.
2. **The measurement units of the variance are always the square of the measurement units of the original data**. This is so because the original values are squared to calculate the variance. In Example 3–12, the measurement units of the original data are millions of dollars. However, the measurement units of the variance are squared millions of dollars, which, of course, does not make any sense. Thus, the variance of the 2010 baggage fee revenue of the six airlines in Example 3–12 is 77,709.06666 squared million dollars. But the measurement units of the standard deviation are the same as the measurement units of the original data because the standard deviation is obtained by taking the square root of the variance.

■ EXAMPLE 3–13

Calculating the population variance and standard deviation for ungrouped data.

Following are the 2011 earnings (in thousands of dollars) before taxes for all six employees of a small company.

88.50 108.40 65.50 52.50 79.80 54.60

Calculate the variance and standard deviation for these data.

Solution Let x denote the 2011 earnings before taxes of an employee of this company. The values of Σx and Σx^2 are calculated in Table 3.7.

Table 3.7

x	x^2
88.50	7832.25
108.40	11,750.56
65.50	4290.25
52.50	2756.25
79.80	6368.04
54.60	2981.16
$\Sigma x = 449.30$	$\Sigma x^2 = 35{,}978.51$

Because the data in this example are on earnings of *all* employees of this company, we use the population formula to compute the variance. Thus, the variance is

$$\sigma^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{N}}{N} = \frac{35{,}978.51 - \frac{(449.30)^2}{6}}{6} = \mathbf{388.90}$$

The standard deviation is obtained by taking the (positive) square root of the variance:

$$\sigma = \sqrt{\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{N}}{N}} = \sqrt{388.90} = \mathbf{19.721\ thousand} = \mathbf{\$19{,}721}$$

Thus, the standard deviation of the 2011 earnings of all six employees of this company is $19,721. ■

Warning ▶ Note that Σx^2 is not the same as $(\Sigma x)^2$. The value of Σx^2 is obtained by squaring the x values and then adding them. The value of $(\Sigma x)^2$ is obtained by squaring the value of Σx.

The uses of the standard deviation are discussed in Section 3.4. Later chapters explain how the mean and the standard deviation taken together can help in making inferences about the population.

3.2.3 Population Parameters and Sample Statistics

A numerical measure such as the mean, median, mode, range, variance, or standard deviation calculated for a population data set is called a *population parameter*, or simply a **parameter**. A summary measure calculated for a sample data set is called a *sample statistic*, or simply a **statistic**. Thus, μ and σ are population parameters, and $\bar{x}$ and s are sample statistics. As an illustration, $\bar{x}$ = $139.5 million in Example 3–1 is a sample statistic, and μ = 45.25 years in Example 3–2 is a population parameter. Similarly, s = $278.76 million in Example 3–12 is a sample statistic, whereas σ = $19,721 in Example 3–13 is a population parameter.

EXERCISES

CONCEPTS AND PROCEDURES

3.37 The range, as a measure of spread, has the disadvantage of being influenced by outliers. Illustrate this with an example.

3.38 Can the standard deviation have a negative value? Explain.

3.39 When is the value of the standard deviation for a data set zero? Give one example. Calculate the standard deviation for the example and show that its value is zero.

3.40 Briefly explain the difference between a population parameter and a sample statistic. Give one example of each.

3.41 The following data set belongs to a population:

5   −7   2   0   −9   16   10   7

Calculate the range, variance, and standard deviation.

3.42 The following data set belongs to a sample:

14   16   −10   8   8   −18

Calculate the range, variance, and standard deviation.

APPLICATIONS

3.43 The following data give the number of shoplifters apprehended during each of the past 10 weeks at a large department store.

7 10 8 3 15 12 6 11 10 8

a. Find the mean for these data. Calculate the deviations of the data values from the mean. Is the sum of these deviations zero?
b. Calculate the range, variance, and standard deviation.

3.44 The following data give the prices of seven textbooks randomly selected from a university bookstore.

$112 $170 $93 $113 $56 $161 $123

a. Find the mean for these data. Calculate the deviations of the data values from the mean. Is the sum of these deviations zero?
b. Calculate the range, variance, and standard deviation.

3.45 Refer to Exercise 3.20, which listed the number of colors of M&Ms that each of the 14 Fun Size bags contained. Those data are reproduced here:

5 6 4 4 6 3 2 4 5 4 3 6 3 6

Calculate the range, variance, and standard deviation.

3.46 Refer to the data in Exercise 3.23, which contained the numbers of tornadoes that touched down in 12 states that had the most tornadoes during the period 1950 to 1994. The data are reproduced here.

1113 2009 1374 1137 2110 1086 1166 1039 1673 2300 1139 5490

Find the range, variance, and standard deviation for these data.

3.47 The following data give the numbers of pieces of junk mail received by 10 families during the past month.

41 18 28 11 29 19 14 31 33 36

Find the range, variance, and standard deviation.

3.48 The following data give the number of highway collisions with large wild animals, such as deer or moose, in one of the northeastern states during each week of a 9-week period.

7 10 5 8 2 6 7 3 9

Find the range, variance, and standard deviation.

3.49 Refer to Exercise 3.24, which listed the number of women from each of 12 countries who were on the Rolex Women's World Golf Rankings Top 50 list as of July 18, 2011. Those data are reproduced here:

3 1 1 1 10 1 1 1 18 2 3 8

Calculate the range, variance, and standard deviation.

3.50 The following data give the number of hot dogs consumed by 10 participants in a hot-dog-eating contest.

21 17 32 5 25 15 17 21 9 24

Calculate the range, variance, and standard deviation for these data.

3.51 Following are the temperatures (in degrees Fahrenheit) observed during eight wintry days in a midwestern city:

23 14 6 −7 −2 11 16 19

Compute the range, variance, and standard deviation.

3.52 Refer to Exercise 2.94, which listed the alcohol content by volume for each of the 13 varieties of beer produced by Sierra Nevada Brewery. Those data are reproduced here:

4.4 5.0 5.0 5.6 5.6 5.8 5.9 5.9 6.7 6.8 6.9 7.0 9.6

Calculate the range, variance, and standard deviation.

3.53 The following data represent the total points scored in each of the NFL Super Bowl games played from 2001 through 2012, in that order:

41 37 69 61 45 31 46 31 50 48 56 38

Compute the range, variance, and standard deviation for these data.

3.54 The following data represent the 2011 guaranteed salaries (in thousands of dollars) of the head coaches of the final eight teams in the 2011 NCAA Men's Basketball Championship. The data represent the 2011 salaries of basketball coaches of the following universities, entered in that order: Arizona, Butler, Connecticut, Florida, Kansas, Kentucky, North Carolina, and Virginia Commonwealth. (*Source:* www.usatoday.com).

1950 434 2300 3575 3376 3800 1655 418

Compute the range, variance, and standard deviation for these data.

3.55 The following data give the hourly wage rates of eight employees of a company.

$22 22 22 22 22 22 22 22

Calculate the standard deviation. Is its value zero? If yes, why?

3.56 The following data are the ages (in years) of six students.

20 20 20 20 20 20

Calculate the standard deviation. Is its value zero? If yes, why?

*3.57 One disadvantage of the standard deviation as a measure of dispersion is that it is a measure of absolute variability and not of relative variability. Sometimes we may need to compare the variability of two different data sets that have different units of measurement. The **coefficient of variation** is one such measure. The coefficient of variation, denoted by CV, expresses standard deviation as a percentage of the mean and is computed as follows:

$$\text{For population data:} \quad \text{CV} = \frac{\sigma}{\mu} \times 100\%$$

$$\text{For sample data:} \quad \text{CV} = \frac{s}{\bar{x}} \times 100\%$$

The yearly salaries of all employees who work for a company have a mean of $62,350 and a standard deviation of $6820. The years of experience for the same employees have a mean of 15 years and a standard deviation of 2 years. Is the relative variation in the salaries larger or smaller than that in years of experience for these employees?

*3.58 The SAT scores of 100 students have a mean of 1020 and a standard deviation of 115. The GPAs of the same 100 students have a mean of 3.21 and a standard deviation of .26. Is the relative variation in SAT scores larger or smaller than that in GPAs?

*3.59 Consider the following two data sets.

Data Set I:	12	25	37	8	41
Data Set II:	19	32	44	15	48

Note that each value of the second data set is obtained by adding 7 to the corresponding value of the first data set. Calculate the standard deviation for each of these two data sets using the formula for sample data. Comment on the relationship between the two standard deviations.

*3.60 Consider the following two data sets.

Data Set I:	4	8	15	9	11
Data Set II:	12	24	45	27	33

Note that each value of the second data set is obtained by multiplying the corresponding value of the first data set by 3. Calculate the standard deviation for each of these two data sets using the formula for population data. Comment on the relationship between the two standard deviations.

3.3 Mean, Variance, and Standard Deviation for Grouped Data

In Sections 3.1.1 and 3.2.2, we learned how to calculate the mean, variance, and standard deviation for ungrouped data. In this section, we will learn how to calculate the mean, variance, and standard deviation for grouped data.

3.3.1 Mean for Grouped Data

We learned in Section 3.1.1 that the mean is obtained by dividing the sum of all values by the number of values in a data set. However, if the data are given in the form of a frequency table, we no longer know the values of individual observations. Consequently, in such cases, we cannot obtain the sum of individual values. We find an approximation for the sum of these values using the procedure explained in the next paragraph and example. The formulas used to calculate the mean for grouped data follow.

Calculating Mean for Grouped Data

Mean for population data: $\mu = \frac{\Sigma mf}{N}$

Mean for sample data: $\bar{x} = \frac{\Sigma mf}{n}$

where m is the midpoint and f is the frequency of a class.

To calculate the mean for grouped data, first find the midpoint of each class and then multiply the midpoints by the frequencies of the corresponding classes. The sum of these products, denoted by Σmf, gives an approximation for the sum of all values. To find the value of the mean, divide this sum by the total number of observations in the data.

■ EXAMPLE 3–14

Calculating the population mean for grouped data.

Table 3.8 gives the frequency distribution of the daily commuting times (in minutes) from home to work for *all* 25 employees of a company.

Table 3.8

Daily Commuting Time (minutes)	Number of Employees
0 to less than 10	4
10 to less than 20	9
20 to less than 30	6
30 to less than 40	4
40 to less than 50	2

Calculate the mean of the daily commuting times.

Solution Note that because the data set includes *all* 25 employees of the company, it represents the population. Table 3.9 shows the calculation of Σmf. Note that in Table 3.9, m denotes the midpoints of the classes.

Table 3.9

Daily Commuting Time (minutes)	f	m	mf
0 to less than 10	4	5	20
10 to less than 20	9	15	135
20 to less than 30	6	25	150
30 to less than 40	4	35	140
40 to less than 50	2	45	90
	$N = 25$		$\Sigma mf = 535$

To calculate the mean, we first find the midpoint of each class. The class midpoints are recorded in the third column of Table 3.9. The products of the midpoints and the corresponding frequencies are listed in the fourth column. The sum of the fourth column values, denoted by Σmf, gives the approximate total daily commuting time (in minutes) for all 25 employees. The mean is obtained by dividing this sum by the total frequency. Therefore,

$$\mu = \frac{\Sigma mf}{N} = \frac{535}{25} = \mathbf{21.40\ minutes}$$

Thus, the employees of this company spend an average of 21.40 minutes a day commuting from home to work. ■

What do the numbers 20, 135, 150, 140, and 90 in the column labeled *mf* in Table 3.9 represent? We know from this table that 4 employees spend 0 to less than 10 minutes commuting per day. If we assume that the time spent commuting by these 4 employees is evenly spread in the interval 0 to less than 10, then the midpoint of this class (which is 5) gives the mean time spent commuting by these 4 employees. Hence, $4 \times 5 = 20$ is the approximate total time (in minutes) spent commuting per day by these 4 employees. Similarly, 9 employees spend 10 to less than 20 minutes commuting per day, and the total time spent commuting by these 9 employees is approximately 135 minutes a day. The other numbers in this column can be interpreted in the same way. Note that these numbers give the approximate commuting times for these employees based on the assumption of an even spread within classes. The total commuting time for all 25 employees is approximately 535 minutes. Consequently, 21.40 minutes is an approximate and not the exact value of the mean. We can find the exact value of the mean only if we know the exact commuting time for each of the 25 employees of the company.

■ EXAMPLE 3–15

Calculating the sample mean for grouped data.

Table 3.10 gives the frequency distribution of the number of orders received each day during the past 50 days at the office of a mail-order company.

Table 3.10

Number of Orders	Number of Days
10–12	4
13–15	12
16–18	20
19–21	14

Calculate the mean.

Solution Because the data set includes only 50 days, it represents a sample. The value of Σmf is calculated in Table 3.11.

Table 3.11

Number of Orders	*f*	*m*	*mf*
10–12	4	11	44
13–15	12	14	168
16–18	20	17	340
19–21	14	20	280
	$n = 50$		$\Sigma mf = 832$

The value of the sample mean is

$$\bar{x} = \frac{\Sigma mf}{n} = \frac{832}{50} = \mathbf{16.64\ orders}$$

Thus, this mail-order company received an average of 16.64 orders per day during these 50 days. ■

3.3.2 Variance and Standard Deviation for Grouped Data

Following are what we will call the *basic formulas* that are used to calculate the population and sample variances for grouped data:

$$\sigma^2 = \frac{\Sigma f(m - \mu)^2}{N} \quad \text{and} \quad s^2 = \frac{\Sigma f(m - \bar{x})^2}{n - 1}$$

where σ^2 is the population variance, s^2 is the sample variance, and m is the midpoint of a class.

In either case, the standard deviation is obtained by taking the positive square root of the variance.

Again, the *short-cut formulas* are more efficient for calculating the variance and standard deviation. Section A3.1.2 of Appendix 3.1 at the end of this chapter shows how to use the basic formulas to calculate the variance and standard deviation for grouped data.

Short-Cut Formulas for the Variance and Standard Deviation for Grouped Data

$$\sigma^2 = \frac{\Sigma m^2 f - \dfrac{(\Sigma mf)^2}{N}}{N} \quad \text{and} \quad s^2 = \frac{\Sigma m^2 f - \dfrac{(\Sigma mf)^2}{n}}{n - 1}$$

where σ^2 is the population variance, s^2 is the sample variance, and m is the midpoint of a class.

The standard deviation is obtained by taking the positive square root of the variance.

Population standard deviation: $\sigma = \sqrt{\dfrac{\Sigma m^2 f - \dfrac{(\Sigma mf)^2}{N}}{N}}$

Sample standard deviation: $s = \sqrt{\dfrac{\Sigma m^2 f - \dfrac{(\Sigma mf)^2}{n}}{n - 1}}$

Examples 3–16 and 3–17 illustrate the use of these formulas to calculate the variance and standard deviation.

■ EXAMPLE 3–16

Calculating the population variance and standard deviation for grouped data.

The following data, reproduced from Table 3.8 of Example 3–14, give the frequency distribution of the daily commuting times (in minutes) from home to work for all 25 employees of a company.

Daily Commuting Time (minutes)	Number of Employees
0 to less than 10	4
10 to less than 20	9
20 to less than 30	6
30 to less than 40	4
40 to less than 50	2

Calculate the variance and standard deviation.

Solution All four steps needed to calculate the variance and standard deviation for grouped data are shown after Table 3.12.

Table 3.12

Daily Commuting Time (minutes)	f	m	mf	m^2f
0 to less than 10	4	5	20	100
10 to less than 20	9	15	135	2025
20 to less than 30	6	25	150	3750
30 to less than 40	4	35	140	4900
40 to less than 50	2	45	90	4050
	$N = 25$		$\Sigma mf = 535$	$\Sigma m^2f = 14{,}825$

Step 1. *Calculate the value of* Σmf.

To calculate the value of Σmf, first find the midpoint m of each class (see the third column in Table 3.12) and then multiply the corresponding class midpoints and class frequencies (see the fourth column). The value of Σmf is obtained by adding these products. Thus,

$$\Sigma mf = 535$$

Step 2. *Find the value of* Σm^2f.

To find the value of Σm^2f, square each m value and multiply this squared value of m by the corresponding frequency (see the fifth column in Table 3.12). The sum of these products (that is, the sum of the fifth column) gives Σm^2f. Hence,

$$\Sigma m^2f = 14{,}825$$

Step 3. *Calculate the variance.*

Because the data set includes all 25 employees of the company, it represents the population. Therefore, we use the formula for the population variance:

$$\sigma^2 = \frac{\Sigma m^2f - \frac{(\Sigma mf)^2}{N}}{N} = \frac{14{,}825 - \frac{(535)^2}{25}}{25} = \frac{3376}{25} = \mathbf{135.04}$$

Step 4. *Calculate the standard deviation.*

To obtain the standard deviation, take the (positive) square root of the variance.

$$\sigma = \sqrt{\sigma^2} = \sqrt{135.04} = \mathbf{11.62\ minutes}$$

Thus, the standard deviation of the daily commuting times for these employees is 11.62 minutes. ■

Note that the values of the variance and standard deviation calculated in Example 3–16 for grouped data are approximations. The exact values of the variance and standard deviation can be obtained only by using the ungrouped data on the daily commuting times of these 25 employees.

■ EXAMPLE 3–17

Calculating the sample variance and standard deviation for grouped data.

The following data, reproduced from Table 3.10 of Example 3–15, give the frequency distribution of the number of orders received each day during the past 50 days at the office of a mail-order company.

Number of Orders	f
10–12	4
13–15	12
16–18	20
19–21	14

Calculate the variance and standard deviation.

Solution All the information required for the calculation of the variance and standard deviation appears in Table 3.13.

Table 3.13

Number of Orders	f	m	mf	m^2f
10–12	4	11	44	484
13–15	12	14	168	2352
16–18	20	17	340	5780
19–21	14	20	280	5600
	$n = 50$		$\Sigma mf = 832$	$\Sigma m^2f = 14{,}216$

Because the data set includes only 50 days, it represents a sample. Hence, we use the sample formulas to calculate the variance and standard deviation. By substituting the values into the formula for the sample variance, we obtain

$$s^2 = \frac{\Sigma m^2 f - \frac{(\Sigma mf)^2}{n}}{n-1} = \frac{14{,}216 - \frac{(832)^2}{50}}{50-1} = \mathbf{7.5820}$$

Hence, the standard deviation is

$$s = \sqrt{s^2} = \sqrt{7.5820} = \mathbf{2.75\ orders}$$

Thus, the standard deviation of the number of orders received at the office of this mail-order company during the past 50 days is 2.75. ■

EXERCISES

CONCEPTS AND PROCEDURES

3.61 Are the values of the mean and standard deviation that are calculated using grouped data exact or approximate values of the mean and standard deviation, respectively? Explain.

3.62 Using the population formulas, calculate the mean, variance, and standard deviation for the following grouped data.

x	2–4	5–7	8–10	11–13	14–16
f	3	10	14	8	5

3.63 Using the sample formulas, find the mean, variance, and standard deviation for the grouped data displayed in the following table.

x	f
0 to less than 4	17
4 to less than 8	23
8 to less than 12	15
12 to less than 16	11
16 to less than 20	8
20 to less than 24	6

APPLICATIONS

3.64 The following table gives the frequency distribution of the amounts of telephone bills for August 2012 for a sample of 50 families.

Amount of Telephone Bill (dollars)	Number of Families
40 to less than 70	8
70 to less than 100	10
100 to less than 130	16
130 to less than 160	11
160 to less than 190	5

Calculate the mean, variance, and standard deviation.

3.65 The following table gives the frequency distribution of the number of hours spent per week playing video games by all 50 students of the eighth grade at a school.

Hours per Week	Number of Students
0 to less than 5	5
5 to less than 10	8
10 to less than 15	14
15 to less than 20	13
20 to less than 25	8
25 to less than 30	2

Find the mean, variance, and standard deviation.

3.66 The following table gives the grouped data on the weights of all 100 babies born at a hospital in 2009.

Weight (pounds)	Number of Babies
3 to less than 5	3
5 to less than 7	35
7 to less than 9	45
9 to less than 11	15
11 to less than 13	2

Find the mean, variance, and standard deviation.

3.67 The following table gives the frequency distribution of the total miles driven during 2012 by 300 car owners.

Miles Driven in 2012 (in thousands)	Number of Car Owners
0 to less than 5	7
5 to less than 10	26
10 to less than 15	59
15 to less than 20	71
20 to less than 25	62
25 to less than 30	39
30 to less than 35	22
35 to less than 40	14

Find the mean, variance, and standard deviation. Give a brief interpretation of the values in the column labeled mf in your table of calculations. What does Σmf represent?

3.68 The following table gives information on the amounts (in dollars) of electric bills for August 2012 for a sample of 50 families.

Amount of Electric Bill (dollars)	Number of Families
0 to less than 60	3
60 to less than 120	17
120 to less than 180	13
180 to less than 240	11
240 to less than 300	6

Find the mean, variance, and standard deviation. Give a brief interpretation of the values in the column labeled mf in your table of calculations. What does Σmf represent?

3.69 For 50 airplanes that arrived late at an airport during a week, the time by which they were late was observed. In the following table, x denotes the time (in minutes) by which an airplane was late, and f denotes the number of airplanes.

x	f
0 to less than 20	14
20 to less than 40	18
40 to less than 60	9
60 to less than 80	5
80 to less than 100	4

Find the mean, variance, and standard deviation.

3.70 The following table gives the frequency distribution of the number of errors committed by a college baseball team in all of the 45 games that it played during the 2011–12 season.

Number of Errors	Number of Games
0	12
1	16
2	8
3	6
4	2
5	1

Find the mean, variance, and standard deviation. (*Hint:* The classes in this example are single valued. These values of classes will be used as values of m in the formulas for the mean, variance, and standard deviation.)

3.71 The following table gives the frequency distribution of the number of hours spent per week on activities that involve sports and/or exercise by a sample of 400 Americans. The numbers are consistent with the summary results from the Bureau of Labor Statistics' American Time Use Survey (www.bls.gov/tus).

Hours per Week	Number of People
0 to less than 3.5	34
3.5 to less than 7.0	92
7.0 to less than 10.5	55
10.5 to less than 14.0	83
14.0 to less than 28.0	121
28.0 to less than 56.0	15

Find the mean, variance, and standard deviation.

3.4 Use of Standard Deviation

By using the mean and standard deviation, we can find the proportion or percentage of the total observations that fall within a given interval about the mean. This section briefly discusses Chebyshev's theorem and the empirical rule, both of which demonstrate this use of the standard deviation.

3.4.1 Chebyshev's Theorem

Chebyshev's theorem gives a lower bound for the area under a curve between two points that are on opposite sides of the mean and at the same distance from the mean.

Definition

Chebyshev's Theorem For any number k greater than 1, at least $(1 - 1/k^2)$ of the data values lie within k standard deviations of the mean.

Figure 3.5 illustrates Chebyshev's theorem.

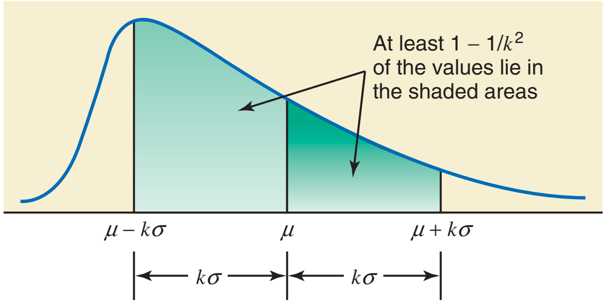


Figure 3.5 Chebyshev's theorem.

Thus, for example, if $k = 2$, then

$$1 - \frac{1}{k^2} = 1 - \frac{1}{(2)^2} = 1 - \frac{1}{4} = 1 - .25 = .75 \text{ or } 75\%$$

Therefore, according to Chebyshev's theorem, at least .75, or 75%, of the values of a data set lie within two standard deviations of the mean. This is shown in Figure 3.6.

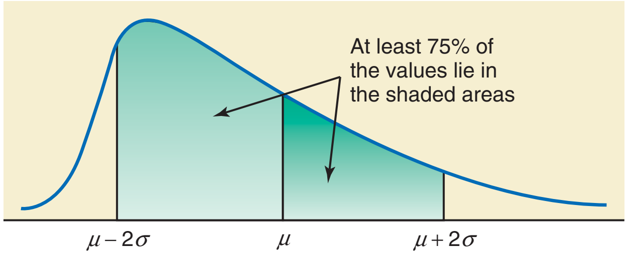


Figure 3.6 Percentage of values within two standard deviations of the mean for Chebyshev's theorem.

If $k = 3$, then,

$$1 - \frac{1}{k^2} = 1 - \frac{1}{(3)^2} = 1 - \frac{1}{9} = 1 - .11 = .89 \text{ or } 89\% \text{ approximately}$$

According to Chebyshev's theorem, at least .89, or 89%, of the values fall within three standard deviations of the mean. This is shown in Figure 3.7.

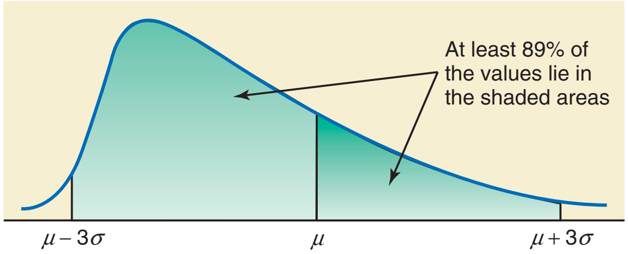


Figure 3.7 Percentage of values within three standard deviations of the mean for Chebyshev's theorem.

Although in Figures 3.5 through 3.7 we have used the population notation for the mean and standard deviation, the theorem applies to both sample and population data. Note that Chebyshev's theorem is applicable to a distribution of any shape. However, Chebyshev's theorem can be used only for $k > 1$. This is so because when $k = 1$, the value of $1 - 1/k^2$ is zero, and when $k < 1$, the value of $1 - 1/k^2$ is negative.

■ EXAMPLE 3–18

Applying Chebyshev's theorem.

The average systolic blood pressure for 4000 women who were screened for high blood pressure was found to be 187 mm Hg with a standard deviation of 22. Using Chebyshev's theorem, find at least what percentage of women in this group have a systolic blood pressure between 143 and 231 mm Hg.

PhotoDisc, Inc./Getty Images

Solution Let μ and σ be the mean and the standard deviation, respectively, of the systolic blood pressures of these women. Then, from the given information,

$$\mu = 187 \quad \text{and} \quad \sigma = 22$$

To find the percentage of women whose systolic blood pressures are between 143 and 231 mm Hg, the first step is to determine k. As shown below, each of the two points, 143 and 231, is 44 units away from the mean.

$$\underset{143}{|} \xleftarrow{} 143 - 187 = -44 \xrightarrow{} \underset{\mu = 187}{|} \xleftarrow{} 231 - 187 = 44 \xrightarrow{} \underset{231}{|}$$

The value of k is obtained by dividing the distance between the mean and each point by the standard deviation. Thus,

$$k = 44/22 = 2$$

$$1 - \frac{1}{k^2} = 1 - \frac{1}{(2)^2} = 1 - \frac{1}{4} = 1 - .25 = .75 \text{ or } \mathbf{75\%}$$

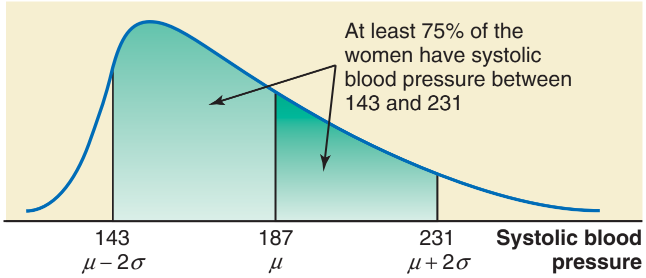


Figure 3.8 Percentage of women with systolic blood pressure between 143 and 231.

Hence, according to Chebyshev's theorem, at least 75% of the women have systolic blood pressure between 143 and 231 mm Hg. This percentage is shown in Figure 3.8. ■

3.4.2 Empirical Rule

Whereas Chebyshev's theorem is applicable to any kind of distribution, the **empirical rule** applies only to a specific type of distribution called a *bell-shaped distribution*, as shown in Figure 3.9. More will be said about such a distribution in Chapter 6, where it is called a *normal curve.* In this section, only the following three rules for the curve are given.

CASE STUDY 3–4

DOES SPREAD MEAN THE SAME AS VARIABILITY AND DISPERSION?

In any discipline, there is terminology that one needs to learn in order to become fluent. Accounting majors need to learn the difference between credits and debits, chemists need to know how an ion differs from an atom, and physical therapists need to know the difference between abduction and adduction. Statistics is no different. Failing to learn the difference between the mean and the median makes much of the remainder of this book very difficult to understand.

Another issue with terminology is the use of words other than the terminology to describe a specific concept or scenario. Sometimes the words one chooses to use can be vague or ambiguous, resulting in confusion. One debate in the statistics community involves the use of the word "spread" in place of the words "dispersion" and "variability." In a 2012 article, "Lexical ambiguity: making a case against *spread*," authors Jennifer Kaplan, Neal Rogness, and Diane Fisher point out that the Oxford English Dictionary has more than 25 definitions for the word spread, many of which students know coming into a statistics class. As a result of knowing some of the meanings of spread, students who use the word spread in place of variability or dispersion "do not demonstrate strong statistical meanings of the word spread at the end of a one-semester statistics course."

In order to examine the extent of this issue, the authors of the article designed a study in which they selected 160 undergraduate students taking an introductory statistics course from 14 different professors at three different universities and in the first week of the semester asked them to write sentences and definitions for spread using its primary meaning. Then, at the end of the semester, the same students were asked to write sentences and definitions for spread using its primary meaning in statistics. The authors found that responses of only one-third of the students related spread to the concept of variability, which has to do with how the data vary around the center of a distribution. A slightly larger percentage of students gave responses that "defined spread as 'to cover evenly or in a thin layer,'" while approximately one in eight responded with a definition that was synonymous with the notion of range. Seven other definitions were given by at least three students in the study.

Although more of the definitions and sentences provided at the end of the course had something to do with statistics, the authors did not see an increase in the percentage of definitions that associated spread with the concept of variability. Hence, they suggested that the ambiguity of the term spread is sufficient enough to stop using it in place of the terms variability and dispersion.

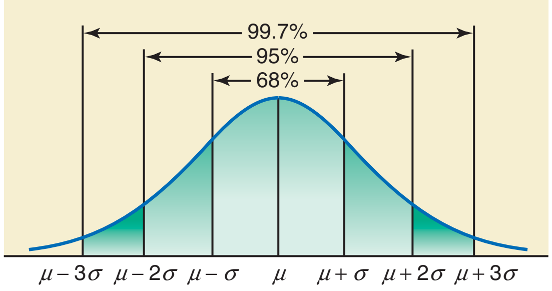
Source: Kaplan, J. J., Rogness, N. T., and Fisher, D. G., "Lexical ambiguity: making a case against *spread*," *Teaching Statistics*, 2011, 34, (2), pp. 56–60. © 2011 Teaching Statistics Trust.

Empirical Rule For a bell-shaped distribution, approximately

1. 68% of the observations lie within one standard deviation of the mean.
2. 95% of the observations lie within two standard deviations of the mean.
3. 99.7% of the observations lie within three standard deviations of the mean.

Figure 3.9 illustrates the empirical rule. Again, the empirical rule applies to both population data and sample data.

Figure 3.9 Illustration of the empirical rule.



■ EXAMPLE 3–19

Applying the empirical rule.

The age distribution of a sample of 5000 persons is bell shaped with a mean of 40 years and a standard deviation of 12 years. Determine the approximate percentage of people who are 16 to 64 years old.

Solution We use the empirical rule to find the required percentage because the distribution of ages follows a bell-shaped curve. From the given information, for this distribution,

$$\bar{x} = 40 \text{ years} \quad \text{and} \quad s = 12 \text{ years}$$

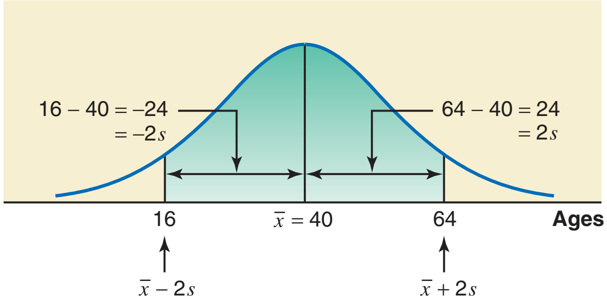


Figure 3.10 Percentage of people who are 16 to 64 years old.

Each of the two points, 16 and 64, is 24 units away from the mean. Dividing 24 by 12, we convert the distance between each of the two points and the mean in terms of standard deviations. Thus, the distance between 16 and 40 and that between 40 and 64 is each equal to $2s$. Consequently, as shown in Figure 3.10, the area from 16 to 64 is the area from $\bar{x} - 2s$ to $\bar{x} + 2s$.

Because the area within two standard deviations of the mean is approximately 95% for a bell-shaped curve, approximately **95%** of the people in the sample are 16 to 64 years old. ■

EXERCISES

■ CONCEPTS AND PROCEDURES

3.72 Briefly explain Chebyshev's theorem and its applications.

3.73 Briefly explain the empirical rule. To what kind of distribution is it applied?

3.74 A sample of 2000 observations has a mean of 74 and a standard deviation of 12. Using Chebyshev's theorem, find at least what percentage of the observations fall in the intervals $\bar{x} \pm 2s$, $\bar{x} \pm 2.5s$, and $\bar{x} \pm 3s$. Note that here $\bar{x} \pm 2s$ represents the interval $\bar{x} - 2s$ to $\bar{x} + 2s$, and so on.

3.75 A large population has a mean of 230 and a standard deviation of 41. Using Chebyshev's theorem, find at least what percentage of the observations fall in the intervals $\mu \pm 2\sigma$, $\mu \pm 2.5\sigma$, and $\mu \pm 3\sigma$.

3.76 A large population has a bell-shaped distribution with a mean of 310 and a standard deviation of 37. Using the empirical rule, find what percentage of the observations fall in the intervals $\mu \pm 1\sigma$, $\mu \pm 2\sigma$, and $\mu \pm 3\sigma$.

3.77 A sample of 3000 observations has a bell-shaped distribution with a mean of 82 and a standard deviation of 16. Using the empirical rule, find what percentage of the observations fall in the intervals $\bar{x} \pm 1s$, $\bar{x} \pm 2s$, and $\bar{x} \pm 3s$.

■ APPLICATIONS

3.78 The mean time taken by all participants to run a road race was found to be 250 minutes with a standard deviation of 30 minutes. Using Chebyshev's theorem, find the percentage of runners who ran this road race in

a. 190 to 310 minutes **b.** 160 to 340 minutes **c.** 175 to 325 minutes

3.79 The 2011 gross sales of all companies in a large city have a mean of \$2.3 million and a standard deviation of \$.5 million. Using Chebyshev's theorem, find at least what percentage of companies in this city had 2009 gross sales of

a. \$1.3 to \$3.3 million **b.** \$1.05 to \$3.55 million **c.** \$.8 to \$3.8 million

3.80 According to the National Center for Education Statistics (www.nces.ed.gov), the amounts of all loans, including Federal Parent PLUS loans, granted to students during the 2007–2008 academic year had a distribution with a mean of \$8109.65. Suppose that the standard deviation of this distribution is \$2412.

a. Using Chebyshev's theorem, find at least what percentage of students had 2007–2008 such loans between

i. \$2079.65 and \$14,139.65 **ii.** \$3285.65 and \$12,933.65

***b.** Using Chebyshev's theorem, find the interval that contains the amounts of 2007–2008 such loans for at least 89% of all students.

3.81 The mean monthly mortgage paid by all home owners in a town is \$2365 with a standard deviation of \$340.

a. Using Chebyshev's theorem, find at least what percentage of all home owners in this town pay a monthly mortgage of

i. \$1685 to \$3045 **ii.** \$1345 to \$3385

***b.** Using Chebyshev's theorem, find the interval that contains the monthly mortgage payments of at least 84% of all home owners in this town.

3.82 The mean life of a certain brand of auto batteries is 44 months with a standard deviation of 3 months. Assume that the lives of all auto batteries of this brand have a bell-shaped distribution. Using the empirical rule, find the percentage of auto batteries of this brand that have a life of

a. 41 to 47 months **b.** 38 to 50 months **c.** 35 to 53 months

3.83 According to the Kaiser Family Foundation, U.S. workers who had employer-provided health insurance paid an average premium of \$4129 for family coverage during 2011 (*USA TODAY*, October 10, 2011). Suppose that the premiums for such family coverage paid this year by all such workers have a bell-shaped distribution with a mean of \$4129 and a standard deviation of \$600. Using the empirical rule, find the approximate percentage of such workers who pay premiums for such family coverage between

a. \$2329 and \$5929 **b.** \$3529 and \$4729 **c.** \$2929 and \$5329

3.84 The prices of all college textbooks follow a bell-shaped distribution with a mean of \$180 and a standard deviation of \$30.

a. Using the empirical rule, find the percentage of all college textbooks with their prices between

i. \$150 and \$210 **ii.** \$120 and \$240

***b.** Using the empirical rule, find the interval that contains the prices of 99.7% of college textbooks.

3.85 Suppose that on a certain section of I-95 with a posted speed limit of 65 mph, the speeds of all vehicles have a bell-shaped distribution with a mean of 72 mph and a standard deviation of 3 mph.

a. Using the empirical rule, find the percentage of vehicles with the following speeds on this section of I-95.

i. 63 to 81 mph **ii.** 69 to 75 mph

***b.** Using the empirical rule, find the interval that contains the speeds of 95% of vehicles traveling on this section of I-95.

3.5 Measures of Position

A **measure of position** determines the position of a single value in relation to other values in a sample or a population data set. There are many measures of position; however, only quartiles, percentiles, and percentile rank are discussed in this section.

3.5.1 Quartiles and Interquartile Range

Quartiles are the summary measures that divide a ranked data set into four equal parts. Three measures will divide any data set into four equal parts. These three measures are the **first quartile** (denoted by Q_1), the **second quartile** (denoted by Q_2), and the **third quartile** (denoted by Q_3). The data should be ranked in increasing order before the quartiles are determined. The quartiles are defined as follows. Note that Q_1 and Q_3 are also called the lower and the upper quartiles, respectively.

Definition

Quartiles *Quartiles* are three summary measures that divide a ranked data set into four equal parts. The second quartile is the same as the median of a data set. The first quartile is the value of the middle term among the observations that are less than the median, and the third quartile is the value of the middle term among the observations that are greater than the median.

Figure 3.11 describes the positions of the three quartiles.

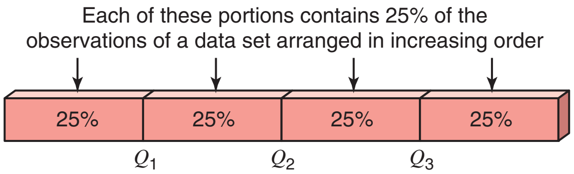

Each of these portions contains 25% of the
observations of a data set arranged in increasing order
25%
25%
25%
25%
Q_1
Q_2
Q_3

Figure 3.11 Quartiles.

Approximately 25% of the values in a ranked data set are less than Q_1 and about 75% are greater than Q_1. The second quartile, Q_2, divides a ranked data set into two equal parts; hence, the second quartile and the median are the same. Approximately 75% of the data values are less than Q_3 and about 25% are greater than Q_3.

The difference between the third quartile and the first quartile for a data set is called the **interquartile range (IQR)**, which is a measure of dispersion.

Calculating Interquartile Range The difference between the third and the first quartiles gives the *interquartile range*; that is,

$$\text{IQR} = \text{Interquartile range} = Q_3 - Q_1$$

Examples 3–20 and 3–21 show the calculation of the quartiles and the interquartile range.

■ EXAMPLE 3–20

Finding quartiles and the interquartile range.

Table 3.3 in Example 3–5 gave the total compensations (in millions of dollars) for the year 2010 of the 12 highest-paid CEOs of U.S. companies. That table is reproduced here:

CEO and Company	2010 Total Compensation (millions of dollars)
Michael D. White (DirecTV)	32.9
David N. Farr (Emerson Electric)	22.9
Brian L. Roberts (Comcast)	28.2
Philippe P. Dauman (Viacom)	84.5
William C. Weldon (Johnson & Johnson)	21.6
Robert A. Iger (Walt Disney)	28.0
Ray R. Iran (Occidental Petroleum)	76.1
Samuel J. Palmisano (IBM)	25.2
John F. Lundgren (Stanley Black & Decker)	32.6
Lawrence J. Ellison (Oracle)	70.1
Alan Mulally (Ford Motor)	26.5
Howard Schultz (Starbucks)	21.7

(a) Find the values of the three quartiles. Where does the total compensation of Michael D. White (CEO of DirecTV) fall in relation to these quartiles?

(b) Find the interquartile range.

Solution

Finding quartiles for an even number of data values.

(a) First we rank the given data in increasing order. Then we calculate the three quartiles as follows:

21.6 21.7 22.9 25.2 26.5 28.0 28.2 32.6 32.9 70.1 76.1 84.5

Values less than the median Values greater than the median

21.6 21.7 22.9 ↑ 25.2 26.5 28.0 ↑ 28.2 32.6 32.9 ↑ 70.1 76.1 84.5

$$Q_1 = \frac{22.9 + 25.2}{2} = \mathbf{24.05} \qquad Q_2 = \frac{28.0 + 28.2}{2} = \mathbf{28.1} \qquad Q_3 = \frac{32.9 + 70.1}{2} = \mathbf{51.5}$$

↑ Also the median

The value of Q_2, which is also the median, is given by the value of the middle term in the ranked data set. For the data of this example, this value is the average of the sixth and seventh terms. Consequently, Q_2 is \$28.1 million. The value of Q_1 is given by the value of the middle term of the six values that fall below the median (or Q_2). Thus, it is obtained by taking the average of the third and fourth terms. So, Q_1 is \$24.05 million. The value of Q_3 is given by the value of the middle term of the six values that fall above the median. For the data of this example, Q_3 is obtained by taking the average of the ninth and tenth terms, and it is \$51.5 million.

The value of $Q_1 =$ \$24.05 million indicates that 25% of the CEOs in this sample had 2010 total compensations less than \$24.05 million and 75% of them had 2010 total compensations higher than \$24.05 million. Similarly, we can state that half of these CEOs had 2010 total compensations less than \$28.1 million and the other half had higher than \$28.1 million, since the second quartile is \$28.1 million. The value of $Q_3 =$ \$51.5 million indicates that 75% of these CEOs had 2010 total compensations less than \$51.5 million and 25% had higher than this value.

By looking at the position of \$32.9 million (total compensation of Michael D. White, CEO of DirecTV), we can state that this value lies in the **bottom 75%** of these 2010 total compensations and is just below Q_3. This value falls between the second and third quartiles.

Finding the interquartile range.

(b) The interquartile range is given by the difference between the values of the third and first quartiles. Thus,

$$\text{IQR} = \text{Interquartile range} = Q_3 - Q_1 = 51.5 - 24.05 = \mathbf{\$27.45\ million}$$ ■

■ EXAMPLE 3–21

Finding quartiles and the interquartile range.

The following are the ages (in years) of nine employees of an insurance company:

47 28 39 51 33 37 59 24 33

(a) Find the values of the three quartiles. Where does the age of 28 years fall in relation to the ages of these employees?

(b) Find the interquartile range.

Solution

Finding quartiles for an odd number of data values.

(a) First we rank the given data in increasing order. Then we calculate the three quartiles as follows:

$$\overbrace{24\qquad 28\qquad 33\qquad 33}^{\text{Values less than the median}}\qquad 37\qquad \overbrace{39\qquad 47\qquad 51\qquad 59}^{\text{Values greater than the median}}$$

$$Q_1 = \frac{28 + 33}{2} = \mathbf{30.5} \qquad Q_2 = \mathbf{37} \qquad Q_3 = \frac{47 + 51}{2} = \mathbf{49}$$

Also the median

Thus the values of the three quartiles are

$$Q_1 = \mathbf{30.5\ years}, \quad Q_2 = \mathbf{37\ years}, \quad \text{and} \quad Q_3 = \mathbf{49\ years}$$

The age of 28 falls in the **lowest 25%** of the ages.

Finding the interquartile range.

(b) The interquartile range is

$$\text{IQR} = \text{Interquartile range} = Q_3 - Q_1 = 49 - 30.5 = \mathbf{18.5\ years}$$ ■

3.5.2 Percentiles and Percentile Rank

Percentiles are the summary measures that divide a ranked data set into 100 equal parts. Each (ranked) data set has 99 percentiles that divide it into 100 equal parts. The data should be ranked in increasing order to compute percentiles. The *k*th percentile is denoted by P_k, where k is an integer in the range 1 to 99. For instance, the 25th percentile is denoted by P_{25}. Figure 3.12 shows the positions of the 99 percentiles.

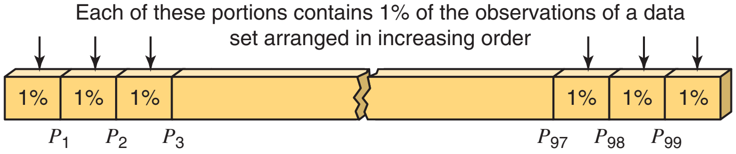

Each of these portions contains 1% of the observations of a data set arranged in increasing order
1%
1%
1%
1%
1%
1%
P_1
P_2
P_3
P_{97}
P_{98}
P_{99}

Figure 3.12 Percentiles.

Thus, the *k*th percentile, P_k, can be defined as a value in a data set such that about k% of the measurements are smaller than the value of P_k and about $(100 - k)$% of the measurements are greater than the value of P_k.

The approximate value of the *k*th percentile is determined as explained next.

Calculating Percentiles The (approximate) value of the *k*th *percentile*, denoted by P_k, is

$$P_k = \text{Value of the } \left(\frac{kn}{100}\right)\text{th term in a ranked data set}$$

where k denotes the number of the percentile and n represents the sample size.

Example 3–22 describes the procedure to calculate the percentiles. For convenience, we will round $kn/100$ to the nearest whole number to find the value of P_k.

■ EXAMPLE 3–22

Finding the percentile for a data set.

Refer to the data on total compensations (in millions of dollars) for the year 2010 of the 12 highest-paid CEOs of U.S. companies given in Exercise 3–20. Find the value of the 60th percentile. Give a brief interpretation of the 60th percentile.

Solution From Example 3–20, the data arranged in increasing order are as follows:

21.6 21.7 22.9 25.2 26.5 28.0 28.2 32.6 32.9 70.1 76.1 84.5

The position of the 60th percentile is

$$\frac{kn}{100} = \frac{60(12)}{100} = 7.20\text{th term} \simeq 7\text{th term}$$

The value of the 7.20th term can be approximated by the value of the 7th term in the ranked data. Therefore,

$$P_{60} = \text{60th percentile} = 28.2 = \mathbf{\$28.2\ million}$$

Thus, approximately 60% of these 12 CEOs had 2010 total compensations less than $28.2 million. ■

We can also calculate the **percentile rank** for a particular value x_i of a data set by using the formula given below. The percentile rank of x_i gives the percentage of values in the data set that are less than x_i.

Finding Percentile Rank of a Value

$$\text{Percentile rank of } x_i = \frac{\text{Number of values less than } x_i}{\text{Total number of values in the data set}} \times 100\%$$

Example 3–23 shows how the percentile rank is calculated for a data value.

Finding the percentile rank for a data value.

■ EXAMPLE 3–23

Refer to the data on total compensations (in millions of dollars) for the year 2010 of the 12 highest-paid CEOs of U.S. companies given in Exercise 3–20. Find the percentile rank for \$26.5 million (2010 total compensation of Alan Mulally, CEO of Ford Motor). Give a brief interpretation of this percentile rank.

Solution From Example 3–20, the data arranged in increasing order are as follows:

21.6 21.7 22.9 25.2 26.5 28.0 28.2 32.6 32.9 70.1 76.1 84.5

In this data set, 4 of the 12 values are less than \$26.5 million. Hence,

$$\text{Percentile rank of } 26.5 = \frac{4}{12} \times 100 = \mathbf{33.33\%}$$

Rounding this answer to the nearest integral value, we can state that about 33% of these 12 CEOs had 2010 total compensations less than \$26.5 million. Hence, 67% of these 12 CEOs had \$26.5 million or higher total compensations in 2010. ■

Most statistical software packages use slightly different methods to calculate quartiles and percentiles. Those methods, while more precise, are beyond the scope of this text.

EXERCISES

■ CONCEPTS AND PROCEDURES

3.86 Briefly describe how the three quartiles are calculated for a data set. Illustrate by calculating the three quartiles for two examples, the first with an odd number of observations and the second with an even number of observations.

3.87 Explain how the interquartile range is calculated. Give one example.

3.88 Briefly describe how the percentiles are calculated for a data set.

3.89 Explain the concept of the percentile rank for an observation of a data set.

■ APPLICATIONS

3.90 The following data give the weights (in pounds) lost by 15 members of a health club at the end of 2 months after joining the club.

4 10 8 7 24 12 5 13
11 10 20 9 8 9 18

a. Compute the values of the three quartiles and the interquartile range.
b. Calculate the (approximate) value of the 82nd percentile.
c. Find the percentile rank of 10.

3.91 The following data give the speeds of 14 cars (in mph) measured by radar, traveling on I-84.

73 75 69 68 78 74 74
76 65 79 69 77 71 72

a. Find the values of the three quartiles and the interquartile range.
b. Calculate the (approximate) value of the 35th percentile.
c. Compute the percentile rank of 71.

3.92 The following data give the numbers of computer keyboards assembled at the Twentieth Century Electronics Company for a sample of 25 days.

45 52 48 41 54 46 44 40 48 53
51 53 48 48 46 43 52 50 54 50
42 47 50 49 52

a. Calculate the values of the three quartiles and the interquartile range.
b. Determine the (approximate) value of the 53rd percentile.
c. Find the percentile rank of 50.

3.93 The following data give the numbers of minor penalties accrued by each of the 30 National Hockey League franchises during the 2010–11 regular season.

249	265	269	287	287	292	299	300	300	301
302	304	311	312	320	325	330	331	335	337
344	347	347	348	352	353	354	355	363	374

a. Calculate the values of the three quartiles and the interquartile range.
b. Find the approximate value of the 57th percentile.
c. Calculate the percentile rank of 311.

3.94 The following data give the numbers of text messages sent by a high school student on 40 randomly selected days during 2012:

32	33	33	34	35	36	37	37	37	32
38	39	40	41	41	42	42	42	43	44
44	45	45	45	47	58	47	47	41	48
48	60	50	50	51	36	53	54	59	61

a. Calculate the values of the three quartiles and the interquartile range. Where does the value 49 fall in relation to these quartiles?
b. Determine the approximate value of the 91st percentile. Give a brief interpretation of this percentile.
c. For what percentage of the days was the number of text messages sent 40 or higher? Answer by finding the percentile rank of 40.

3.95 Nixon Corporation manufactures computer monitors. The following data give the numbers of computer monitors produced at the company for a sample of 30 days.

24	32	27	23	33	33	29	25	23	36
26	26	31	20	27	33	27	23	28	29
31	35	34	22	37	28	23	35	31	43

a. Calculate the values of the three quartiles and the interquartile range. Where does the value of 31 lie in relation to these quartiles?
b. Find the (approximate) value of the 65th percentile. Give a brief interpretation of this percentile.
c. For what percentage of the days was the number of computer monitors produced 32 or higher? Answer by finding the percentile rank of 32.

3.96 The following data give the numbers of new cars sold at a dealership during a 20-day period.

8	5	12	3	9	10	6	12	8	5
4	16	10	14	7	7	3	2	9	11

a. Calculate the values of the three quartiles and the interquartile range. Where does the value of 4 lie in relation to these quartiles?
b. Find the (approximate) value of the 25th percentile. Give a brief interpretation of this percentile.
c. Find the percentile rank of 10. Give a brief interpretation of this percentile rank.

3.97 According to www.money-zine.com, the average FICO score in the United States was around 692 in December 2011. Suppose the following data represent the credit scores of 22 randomly selected loan applicants.

494	728	468	533	747	639	430	690	604	422	356
805	749	600	797	702	628	625	617	647	772	572

a. Calculate the values of the three quartiles and the interquartile range. Where does the value 617 fall in relation to these quartiles?
b. Find the approximate value of the 30th percentile. Give a brief interpretation of this percentile.
c. Calculate the percentile rank of 533. Give a brief interpretation of this percentile rank.

3.6 Box-and-Whisker Plot

A **box-and-whisker plot** gives a graphic presentation of data using five measures: the median, the first quartile, the third quartile, and the smallest and the largest values in the data set between the lower and the upper inner fences. (The inner fences are explained in Example 3–24.) A box-and-whisker plot can help us visualize the center, the spread, and the skewness of a data set. It also helps detect outliers. We can compare different distributions by making box-and-whisker plots for each of them.