CHAPTER 4

Hideji Watanabe/Sebun Photo/
amana images/Getty Images

Numerical Methods for Describing Data

In 2006, Medicare introduced a new prescription drug program. The article "Those Most in Need May Miss Drug Benefit Sign-Up" (*USA Today,* May 9, 2006) notes that only 24% of those eligible for low-income subsidies under this program had signed up just 2 weeks before the enrollment deadline. The article also gave the percentage of those eligible who had signed up in each of 49 states and the District of Columbia (information was not available for Vermont):

24 27 12 38 21 26 23 33 19 19 26 28
16 21 28 20 21 41 22 16 29 26 22 16
27 22 19 22 22 22 30 20 21 34 26 20
25 19 17 21 27 19 27 34 20 30 20 21
14 18

Make the most of your study time by accessing everything you need to succeed online with CourseMate.
Visit http://www.cengagebrain.com where you will find:

- An interactive eBook, which allows you to take notes, highlight, bookmark, search the text, and use in-context glossary definitions
- Step-by-step instructions for Minitab, Excel, TI-83/84, SPSS, and JMP
- Video solutions to selected exercises
- Data sets available for selected examples and exercises
- Online quizzes
- Flashcards
- Videos

What is a typical value for this data set? Is the nationwide figure of 24% representative of the individual state percentages? The enrollment percentages differ widely from state to state, ranging from a low of 12% (Arizona) to a high of 41% (Kentucky). How might we summarize this variability numerically? In this chapter, we show how to calculate numerical summary measures that describe more precisely both the center and the extent of spread in a data set. In Section 4.1, we introduce the mean and the median, the two most widely used measures of the center of a distribution. The variance and the standard deviation are presented in Section 4.2 as measures of variability. In later sections, we will see some additional ways that measures of center and spread can be used to describe data distributions.

4.1 Describing the Center of a Data Set

When describing numerical data, it is common to report a value that is representative of the observations. Such a number describes roughly where the data are located or "centered" along the number line, and is called a measure of center. The two most widely used measures of center are the *mean* and the *median.*

The Mean

The mean of a numerical data set is just the familiar arithmetic average: the sum of the observations divided by the number of observations. It is helpful to have concise notation for the variable on which observations were made, for the number of observations in the data set, and for the individual observations:

$$\begin{aligned} x &= \text{the variable for which we have sample data} \\ n &= \text{the number of observations in the data set (the sample size)} \\ x_1 &= \text{the first observation in the data set} \\ x_2 &= \text{the second observation in the data set} \\ &\vdots \\ x_n &= \text{the } n\text{th (last) observation in the data set} \end{aligned}$$

For example, we might have a sample consisting of $n = 4$ observations on x = battery lifetime (in hours):

$$x_1 = 5.9 \qquad x_2 = 7.3 \qquad x_3 = 6.6 \qquad x_4 = 5.7$$

Notice that the value of the subscript on x has no relationship to the magnitude of the observation. In this example, x_1 is just the first observation in the data set and not necessarily the smallest observation, and x_n is the last observation but not necessarily the largest.

The sum of $x_1, x_2, \ldots, x_n$ can be denoted by $x_1 + x_2 + \cdots + x_n$, but this is cumbersome. The Greek letter $\sum$ is traditionally used in mathematics to denote summation. In particular, $\sum x$ denotes the sum of all the x values in the data set under consideration.*

DEFINITION

The **sample mean** of a sample consisting of numerical observations $x_1, x_2, \ldots,$ x_n, denoted by $\bar{x}$, is

$$\bar{x} = \frac{\text{sum of all observations in the sample}}{\text{number of observations in the sample}} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x}{n}$$

*It is also common to see $\sum x$ written as $\sum x_i$ or even as $\sum_{i=1}^{n} x_i$, but for simplicity we will usually omit the summation indices.

EXAMPLE 4.1 Improving Knee Extension

● Increasing joint extension is one goal of athletic trainers. In a study to investigate the effect of a therapy that uses ultrasound and stretching (Trae Tashiro, Masters Thesis, University of Virginia, 2004) passive knee extension was measured after treatment. Passive knee extension (in degrees) is given for each of 10 participants in the study:

$$\begin{array}{lllll} x_1 = 59 & x_2 = 46 & x_3 = 64 & x_4 = 49 & x_5 = 56 \\ x_6 = 70 & x_7 = 45 & x_8 = 52 & x_9 = 63 & x_{10} = 52 \end{array}$$

The sum of these sample values is $59 + 46 + 64 + \cdots + 52 = 556$, and the sample mean passive knee extension is

$$\bar{x} = \frac{\sum x}{n} = \frac{556}{10} = 55.6$$

We would report 55.6 degrees as a representative value of passive knee extension for this sample (even though there is no person in the sample that actually had a passive knee extension of 55.6 degrees).

The data values in Example 4.1 were all integers, yet the mean was given as 55.6. It is common to use more digits of decimal accuracy for the mean. This allows the value of the mean to fall between possible observable values (for example, the average number of children per family could be 1.8, whereas no single family will have 1.8 children).

The sample mean $\bar{x}$ is computed from sample observations, so it is a characteristic of the particular sample in hand. It is customary to use Roman letters to denote sample characteristics, as we have done with $\bar{x}$. Characteristics of the population are usually denoted by Greek letters. One of the most important of such characteristics is the population mean.

DEFINITION

The **population mean**, denoted by μ, is the average of all x values in the entire population.

For example, the average fuel efficiency for *all* 600,000 cars of a certain type under specified conditions might be $\mu = 27.5$ mpg. A sample of $n = 5$ cars might yield efficiencies of 27.3, 26.2, 28.4, 27.9, 26.5, from which we obtain $\bar{x} = 27.26$ for this particular sample (somewhat smaller than μ). However, a second sample might give $\bar{x} = 28.52$, a third $\bar{x} = 26.85$, and so on. The value of $\bar{x}$ varies from sample to sample, whereas there is just one value for μ. In later chapters, we will see how the value of $\bar{x}$ from a particular sample can be used to draw various conclusions about the value of μ. Example 4.2 illustrates how the value of $\bar{x}$ from a particular sample can differ from the value of μ and how the value of $\bar{x}$ differs from sample to sample.

● Data set available online

EXAMPLE 4.2 County Population Sizes

The 50 states plus the District of Columbia contain 3137 counties. Let x denote the number of residents of a county. Then there are 3137 values of the variable x in the population. The sum of these 3137 values is 293,655,404 (2004 Census Bureau estimate), so the population average value of x is

$$\mu = \frac{293{,}655{,}404}{3137} = 93{,}610.27 \text{ residents per county}$$

We used the Census Bureau web site to select three different samples at random from this population of counties, with each sample consisting of five counties. The results appear in Table 4.1, along with the sample mean for each sample. Not only are the three $\bar{x}$ values different from one another—because they are based on three different samples and the value of $\bar{x}$ depends on the x values in the sample—but also none of the three values comes close to the value of the population mean, μ. If we did not know the value of μ but had only Sample 1 available, we might use $\bar{x}$ as an *estimate* of μ, but our estimate would be far off the mark.

TABLE 4.1 Three Samples from the Population of All U.S. Counties (x = number of residents)

SAMPLE 1		SAMPLE 2		SAMPLE 3	
County	x Value	County	x Value	County	x Value
Fayette, TX	22,513	Stoddard, MO	29,773	Chattahoochee, GA	13,506
Monroe, IN	121,013	Johnston, OK	10,440	Petroleum, MT	492
Greene, NC	20,219	Sumter, AL	14,141	Armstrong, PA	71,395
Shoshone, ID	12,827	Milwaukee, WI	928,018	Smith, MI	14,306
Jasper, IN	31,624	Albany, WY	31,473	Benton, MO	18,519
	$\sum x = 208{,}196$		$\sum x = 1{,}013{,}845$		$\sum x = 118{,}218$
	$\bar{x} = 41{,}639.2$		$\bar{x} = 202{,}769.0$		$\bar{x} = 23{,}643.6$

Alternatively, we could combine the three samples into a single sample with $n = 15$ observations:

$$x_1 = 22{,}513, \ldots, x_5 = 31{,}624, \ldots, x_{15} = 18{,}519$$

$$\sum x = 1{,}340{,}259$$

$$\bar{x} = \frac{1{,}340{,}259}{15} = 89{,}350.6$$

This value is closer to the value of μ but is still somewhat unsatisfactory as an estimate. The problem here is that the population of x values exhibits a lot of variability (the largest value is $x = 9{,}937{,}739$ for Los Angeles County, California, and the smallest value is $x = 52$ for Loving County, Texas, which evidently few people love). Therefore, it is difficult for a sample of 15 observations, let alone just 5, to be reasonably representative of the population. In Chapter 9, you will see how to take variability into account when deciding on a sample size.

One potential drawback to the mean as a measure of center for a data set is that its value can be greatly affected by the presence of even a single *outlier* (an unusually large or small observation) in the data set.

EXAMPLE 4.3 Number of Visits to a Class Web Site

● Forty students were enrolled in a section of a general education course in statistical reasoning during one fall quarter at Cal Poly, San Luis Obispo. The instructor made course materials, grades, and lecture notes available to students on a class web site, and course management software kept track of how often each student accessed any of the web pages on the class site. One month after the course began, the instructor requested a report that indicated how many times each student had accessed a web page on the class site. The 40 observations were:

20	37	4	20	0	84	14	36	5	331	19	0
0	22	3	13	14	36	4	0	18	8	0	26
4	0	5	23	19	7	12	8	13	16	21	7
13	12	8	42								

The sample mean for this data set is $\bar{x} = 23.10$. Figure 4.1 is a Minitab dotplot of the data. Many would argue that 23.10 is not a very representative value for this sample, because 23.10 is larger than most of the observations in the data set—only 7 of 40 observations, or 17.5%, are larger than 23.10. The two outlying values of 84 and 331 (no, that was *not* a typo!) have a substantial impact on the value of $\bar{x}$.

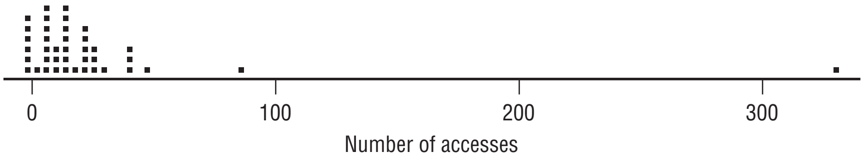


FIGURE 4.1 A Minitab dotplot of the data in Example 4.3.

We now turn our attention to a measure of center that is not as sensitive to outliers—the median.

The Median

The median strip of a highway divides the highway in half, and the median of a numerical data set does the same thing for a data set. Once the data values have been listed in order from smallest to largest, the **median** is the middle value in the list, and it divides the list into two equal parts. Depending on whether the sample size n is even or odd, the process of determining the median is slightly different. When n is an odd number (say, 5), the sample median is the single middle value. But when n is even (say, 6), there are two middle values in the ordered list, and we average these two middle values to obtain the sample median.

Step-by-Step technology instructions available online

● Data set available online

DEFINITION

The **sample median** is obtained by first ordering the n observations from smallest to largest (with any repeated values included, so that every sample observation appears in the ordered list). Then

$$\text{sample median} = \begin{cases} \text{the single middle value if } n \text{ is odd} \\ \text{the average of the middle two values if } n \text{ is even} \end{cases}$$

EXAMPLE 4.4 Web Site Data Revised

The sample size for the web site access data of Example 4.3 was $n = 40$, an even number. The median is the average of the 20th and 21st values (the middle two) in the ordered list of the data. Arranging the data in order from smallest to largest produces the following ordered list (with the two middle values highlighted):

0	0	0	0	0	0	3	4	4	4	5	5
7	7	8	8	8	12	12	13	13	13	14	14
16	18	19	19	20	20	21	22	23	26	36	36
37	42	84	331								

The median can now be determined:

$$median = \frac{13 + 13}{2} = 13$$

Looking at the dotplot (Figure 4.1), we see that this value appears to be a more typical value for the data set than the sample mean $\bar{x} = 23.10$ is.

The sample mean can be sensitive to even a single value that lies far above or below the rest of the data. The value of the mean is pulled out toward such an outlying value or values. The median, on the other hand, is quite *in*sensitive to outliers. For example, the largest sample observation (331) in Example 4.4 can be increased by any amount without changing the value of the median. Similarly, an increase in the second or third largest observations does not affect the median, nor would a decrease in several of the smallest observations.

This stability of the median is what sometimes justifies its use as a measure of center in some situations. For example, the article "Educating Undergraduates on Using Credit Cards" (Nellie Mae, 2005) reported that the mean credit card debt for undergraduate students in 2001 was \$2327, whereas the median credit card debt was only \$1770. In this case, the small percentage of students with unusually high credit card debt may be resulting in a mean that is not representative of a typical student's credit card debt.

Comparing the Mean and the Median

Figure 4.2 shows several smoothed histograms that might represent either a distribution of sample values or a population distribution. Pictorially, the median is the value on the measurement axis that separates the smoothed histogram into two parts, with .5 (50%) of the area under each part of the curve. The mean is a bit harder to visualize. If the histogram were balanced on a triangle (a fulcrum), it would tilt unless the triangle was positioned at the mean. The mean is the balance point for the distribution.

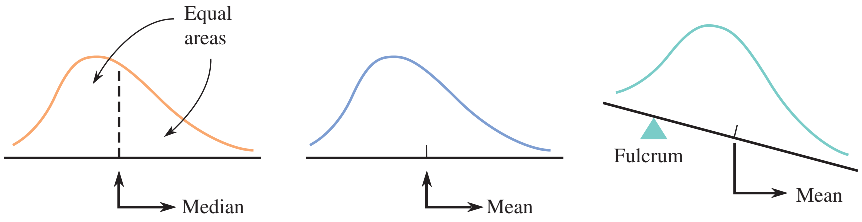


FIGURE 4.2
The mean and the median.

When the histogram is symmetric, the point of symmetry is both the dividing point for equal areas and the balance point, and the mean and the median are equal. However, when the histogram is unimodal (single-peaked) with a longer upper tail (positively skewed), the outlying values in the upper tail pull the mean up, so it generally lies above the median. For example, an unusually high exam score raises the mean but does not affect the median. Similarly, when a unimodal histogram is negatively skewed, the mean is generally smaller than the median (see Figure 4.3).

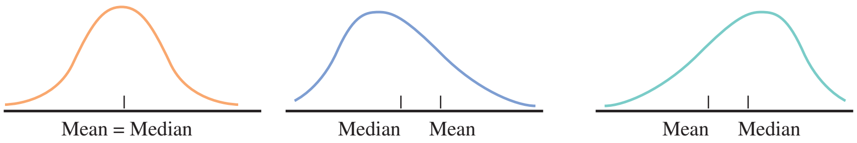


FIGURE 4.3
Relationship between the mean and the median.

Trimmed Means

The extreme sensitivity of the mean to even a single outlier and the extreme insensitivity of the median to a substantial proportion of outliers can sometimes make both of them suspect as a measure of center. A *trimmed mean* is a compromise between these two extremes.

DEFINITION

A **trimmed mean** is computed by first ordering the data values from smallest to largest, deleting a selected number of values from each end of the ordered list, and finally averaging the remaining values.

The **trimming percentage** is the percentage of values deleted from *each* end of the ordered list.

Sometimes the number of observations to be deleted from each end of the data set is specified. Then the corresponding trimming percentage is calculated as

$$\text{trimming percentage} = \left(\frac{\text{number deleted from each end}}{n}\right) \cdot 100$$

In other cases, the trimming percentage is specified and then used to determine how many observations to delete from each end, with

$$\text{number deleted from each end} = \left(\frac{\text{trimming percentage}}{100}\right) \cdot n$$

If the number of observations to be deleted from each end resulting from this calculation is not an integer, it can be rounded to the nearest integer (which changes the trimming percentage a bit).

EXAMPLE 4.5 NBA Salaries

● The web site HoopsHype (hoopshype.com/salaries) publishes salaries of NBA players. Salaries for the players of the Chicago Bulls in 2009 were

Player	2009 Salary
Brad Miller	$12,250,000
Luol Deng	$10,370,425
Kirk Hinrich	$9,500,000
Jerome James	$6,600,000
Tim Thomas	$6,466,600
John Salmons	$5,456,000
Derrick Rose	$5,184,480
Tyrus Thomas	$4,743,598
Joakim Noah	$2,455,680
Jannero Pargo	$2,000,000
James Johnson	$1,594,080
Lindsey Hunter	$1,306,455
Taj Gibson	$1,039,800
Aaron Gray	$1,000,497

A Minitab dotplot of these data is shown in Figure 4.4(a). Because the data distribution is not symmetric and there are outliers, a trimmed mean is a reasonable choice for describing the center of this data set.

There are 14 observations in this data set. Deleting the two largest and the two smallest observations from the data set and then averaging the remaining values would result in a $\left(\frac{2}{14}\right)(100) = 14\%$ trimmed mean. Based on the Bulls' salary data, the two largest salaries are $12,250,000 and $10,370,425, and the two smallest are $1,039,800 and $1,000,497. The average of the remaining 10 observations is

$$\text{14\% trimmed mean} = \frac{9{,}500{,}000 + \cdots + 1{,}306{,}445}{10} = \frac{45{,}306{,}893}{10} = 4{,}530{,}689$$

The mean ($4,997,687) is larger than the trimmed mean because of the few unusually large values in the data set.

For the L.A. Lakers, the difference between the mean ($7,035,947) and the 14% trimmed mean ($5,552,607) is even more dramatic because in 2009 one

● Data set available online

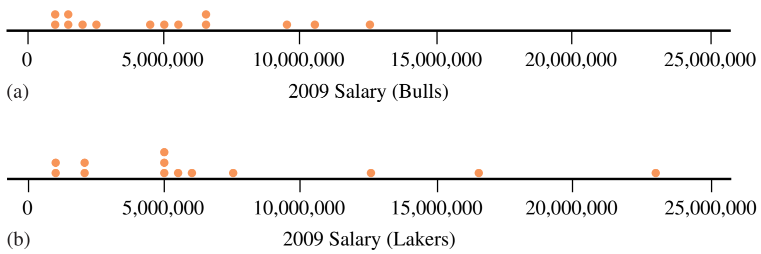


FIGURE 4.4
Minitab dotplots for NBA salary data (a) Bulls (b) Lakers.

player on the Lakers earned over \$23 million and two players earned well over \$10 million (see Figure 4.4(b)).

Categorical Data

The natural numerical summary quantities for a categorical data set are the relative frequencies for the various categories. Each relative frequency is the proportion (fraction) of responses that is in the corresponding category. Often there are only two possible responses (a **dichotomy**)—for example, male or female, does or does not have a driver's license, did or did not vote in the last election. It is convenient in such situations to label one of the two possible responses S (for success) and the other F (for failure). As long as further analysis is consistent with the labeling, it does not matter which category is assigned the S label. When the data set is a sample, the fraction of S's in the sample is called the **sample proportion of successes**.

DEFINITION

The **sample proportion of successes,** denoted by $\hat{p}$, is

$$\hat{p} = \text{sample proportion of successes} = \frac{\text{number of S's in the sample}}{n}$$

where S is the label used for the response designated as success.

EXAMPLE 4.6 Can You Hear Me Now?

Getty Images

It is not uncommon for a cell phone user to complain about the quality of his or her service provider. Suppose that each person in a sample of $n = 15$ cell phone users is asked if he or she is satisfied with the cell phone service. Each response is classified as S (satisfied) or F (not satisfied). The resulting data are

S	F	S	S	S	F	F	S	S	F
S	S	S	F	F					

This sample contains nine S's, so

$$\hat{p} = \frac{9}{15} = .60$$

That is, 60% of the sample responses are S's. Of those surveyed, 60% are satisfied with their cell phone service.

The letter p is used to denote the **population proportion of S's**.* We will see later how the value of $\hat{p}$ from a particular sample can be used to make inferences about p.

*Note that this is one situation in which we will not use a Greek letter to denote a population characteristic. Some statistics books use the symbol π for the population proportion and p for the sample proportion. We will not use π in this context so there is no confusion with the mathematical constant $\pi = 3.14. . . .$

EXERCISES 4.1 - 4.16

4.1 ● The Insurance Institute for Highway Safety (www.iihs.org, June 11, 2009) published data on repair costs for cars involved in different types of accidents. In one study, seven different 2009 models of mini- and micro-cars were driven at 6 mph straight into a fixed barrier. The following table gives the cost of repairing damage to the bumper for each of the seven models.

Model	Repair Cost
Smart Fortwo	\$1,480
Chevrolet Aveo	\$1,071
Mini Cooper	\$2,291
Toyota Yaris	\$1,688
Honda Fit	\$1,124
Hyundai Accent	\$3,476
Kia Rio	\$3,701

Compute the values of the mean and median. Why are these values so different? Which of the two—mean or median—appears to be better as a description of a typical value for this data set?

4.2 ● The article "Caffeinated Energy Drinks—A Growing Problem" (*Drug and Alcohol Dependence* [2009]: 1–10) gave the following data on caffeine concentration (mg/ounce) for eight top-selling energy drinks:

Energy Drink	Caffeine Concentration (mg/oz)
Red Bull	9.6
Monster	10.0
Rockstar	10.0
Full Throttle	9.0
No Fear	10.9
Amp	8.9
SoBe Adrenaline Rush	9.5
Tab Energy	9.1

a. What is the value of the mean caffeine concentration for this set of top-selling energy drinks?

b. Coca-Cola has 2.9 mg/ounce of caffeine and Pepsi Cola has 3.2 mg/ounce of caffeine. Write a sentence explaining how the caffeine concentration of top-selling energy drinks compares to that of these colas.

4.3 ● Consumer Reports Health (www.consumerreports.org/health) reported the accompanying caffeine concentration (mg/cup) for 12 brands of coffee:

Coffee Brand	Caffeine Concentration (mg/cup)
Eight O'Clock	140
Caribou	195
Kickapoo	155
Starbucks	115
Bucks Country Coffee Co.	195
Archer Farms	180
Gloria Jean's Coffees	110
Chock Full o'Nuts	110
Peet's Coffee	130
Maxwell House	55
Folgers	60
Millstone	60

Use at least one measure of center to compare caffeine concentration for coffee with that of the energy drinks of the previous exercise. (Note: 1 cup = 8 ounces)

4.4 ● Consumer Reports Health (www.consumerreports.org/health) reported the sodium content (mg) per 2 tablespoon serving for each of 11 different peanut butters:

120	50	140	120	150	150	150	65
170	250	110					

a. Display these data using a dotplot. Comment on any unusual features of the plot.

b. Compute the mean and median sodium content for the peanut butters in this sample.

c. The values of the mean and the median for this data set are similar. What aspect of the distribution of sodium content—as pictured in the dotplot from Part (a)—provides an explanation for why the values of the mean and median are similar?

4.5 In August 2009, Harris Interactive released the results of the "Great Schools" survey. In this survey, 1086 parents of children attending a public or private school were asked approximately how much time they spent volunteering at school per month over the last school year. For this sample, the mean number of hours per month was 5.6 hours and the median number of hours was 1.0. What does the large difference between the mean and median tell you about this data set?

Bold exercises answered in back ● Data set available online ✦ Video Solution available

4.6 ● The accompanying data on number of minutes used for cell phone calls in one month was generated to be consistent with summary statistics published in a report of a marketing study of San Diego residents (TeleTruth, March 2009):

189	0	189	177	106	201	0	212	0	306
0	0	59	224	0	189	142	83	71	165
236	0	142	236	130					

a. Would you recommend the mean or the median as a measure of center for this data set? Give a brief explanation of your choice. (Hint: It may help to look at a graphical display of the data.)
b. Compute a trimmed mean by deleting the three smallest observations and the three largest observations in the data set and then averaging the remaining 19 observations. What is the trimming percentage for this trimmed mean?
c. What trimming percentage would you need to use in order to delete all of the 0 minute values from the data set? Would you recommend a trimmed mean with this trimming percentage? Explain why or why not.

4.7 ● *USA Today* (May 9, 2006) published the accompanying average weekday circulation for the 6-month period ending March 31, 2006, for the top 20 newspapers in the country:

2,272,815	2,049,786	1,142,464	851,832	724,242
708,477	673,379	579,079	513,387	438,722
427,771	398,329	398,246	397,288	365,011
362,964	350,457	345,861	343,163	323,031

a. Do you think the mean or the median will be larger for this data set? Explain.
b. Compute the values of the mean and the median of this data set.
c. Of the mean and median, which does the best job of describing a typical value for this data set?
d. Explain why it would not be reasonable to generalize from this sample of 20 newspapers to the population of all daily newspapers in the United States.

4.8 ● The chapter introduction gave the accompanying data on the percentage of those eligible for a low-income subsidy who had signed up for a Medicare drug plan in each of 49 states (information was not available for Vermont) and the District of Columbia (*USA Today*, May 9, 2006).

24	27	12	38	21	26	23	33
19	19	26	28	16	21	28	20
21	41	22	16	29	26	22	16
27	22	19	22	22	22	30	20
21	34	26	20	25	19	17	21
27	19	27	34	20	30	20	21
14	18						

a. Compute the mean for this data set.
b. The article stated that nationwide, 24% of those eligible had signed up. Explain why the mean of this data set from Part (a) is not equal to 24. (No information was available for Vermont, but that is not the reason that the mean differs—the 24% was calculated excluding Vermont.)

4.9 ● The U.S. Department of Transportation reported the number of speeding-related crash fatalities for the 20 days of the year that had the highest number of these fatalities between 1994 and 2003 (*Traffic Safety Facts*, July 2005).

Date	Speeding-Related Fatalities	Date	Speeding-Related Fatalities
Jan 1	521	Aug 17	446
Jul 4	519	Dec 24	436
Aug 12	466	Aug 25	433
Nov 23	461	Sep 2	433
Jul 3	458	Aug 6	431
Dec 26	455	Aug 10	426
Aug 4	455	Sept 21	424
Aug 31	446	Jul 27	422
May 25	446	Sep 14	422
Dec 23	446	May 27	420

a. Compute the mean number of speeding-related fatalities for these 20 days.
b. Compute the median number of speeding-related fatalities for these 20 days.
c. Explain why it is not reasonable to generalize from this sample of 20 days to the other 345 days of the year.

4.10 The ministry of Health and Long-Term Care in Ontario, Canada, publishes information on its web site (www.health.gov.on.ca) on the time that patients must wait for various medical procedures. For two cardiac procedures completed in fall of 2005, the following information was provided:

	Number of Completed Procedures	Median Wait Time (days)	Mean Wait Time (days)	90% Completed Within (days)
Angioplasty	847	14	18	39
Bypass surgery	539	13	19	42

a. The median wait time for angioplasty is greater than the median wait time for bypass surgery but the mean wait time is shorter for angioplasty than for bypass surgery. What does this suggest about the distribution of wait times for these two procedures?

b. Is it possible that another medical procedure might have a median wait time that is greater than the time reported for "90% completed within"? Explain.

4.11 Houses in California are expensive, especially on the Central Coast where the air is clear, the ocean is blue, and the scenery is stunning. The median home price in San Luis Obispo County reached a new high in July 2004, soaring to \$452,272 from \$387,120 in March 2004. (*San Luis Obispo Tribune,* April 28, 2004). The article included two quotes from people attempting to explain why the median price had increased. Richard Watkins, chairman of the Central Coast Regional Multiple Listing Services was quoted as saying, "There have been some fairly expensive houses selling, which pulls the median up." Robert Kleinhenz, deputy chief economist for the California Association of Realtors explained the volatility of house prices by stating: "Fewer sales means a relatively small number of very high or very low home prices can more easily skew medians." Are either of these statements correct? For each statement that is incorrect, explain why it is incorrect and propose a new wording that would correct any errors in the statement.

4.12 Consider the following statement: More than 65% of the residents of Los Angeles earn less than the average wage for that city. Could this statement be correct? If so, how? If not, why not?

4.13 ✦ A sample consisting of four pieces of luggage was selected from among those checked at an airline counter, yielding the following data on x = weight (in pounds):

$$x_1 = 33.5, x_2 = 27.3, x_3 = 36.7, x_4 = 30.5$$

Suppose that one more piece is selected and denote its weight by x_5. Find a value of x_5 such that $\bar{x}$ = sample median.

4.14 Suppose that 10 patients with meningitis received treatment with large doses of penicillin. Three days later, temperatures were recorded, and the treatment was considered successful if there had been a reduction in a patient's temperature. Denoting success by S and failure by F, the 10 observations are

S S F S S S F F S S

a. What is the value of the sample proportion of successes?

b. Replace each S with a 1 and each F with a 0. Then calculate $\bar{x}$ for this numerically coded sample. How does $\bar{x}$ compare to $\hat{p}$?

c. Suppose that it is decided to include 15 more patients in the study. How many of these would have to be S's to give $\hat{p}$ = .80 for the entire sample of 25 patients?

4.15 An experiment to study the lifetime (in hours) for a certain brand of light bulb involved putting 10 light bulbs into operation and observing them for 1000 hours. Eight of the light bulbs failed during that period, and those lifetimes were recorded. The lifetimes of the two light bulbs still functioning after 1000 hours are recorded as 1000+. The resulting sample observations were

480 790 1000+ 350 920 860 570 1000+
170 290

Which of the measures of center discussed in this section can be calculated, and what are the values of those measures?

4.16 An instructor has graded 19 exam papers submitted by students in a class of 20 students, and the average so far is 70. (The maximum possible score is 100.) How high would the score on the last paper have to be to raise the class average by 1 point? By 2 points?

4.2 Describing Variability in a Data Set

Reporting a measure of center gives only partial information about a data set. It is also important to describe how much the observations differ from one another. The three different samples displayed in Figure 4.5 all have mean = median = 45. There is a lot of variability in the first sample compared to the third sample. The second sample shows less variability than the first and more variability than the third; most of the variability in the second sample is due to the two extreme values being so far from the center.

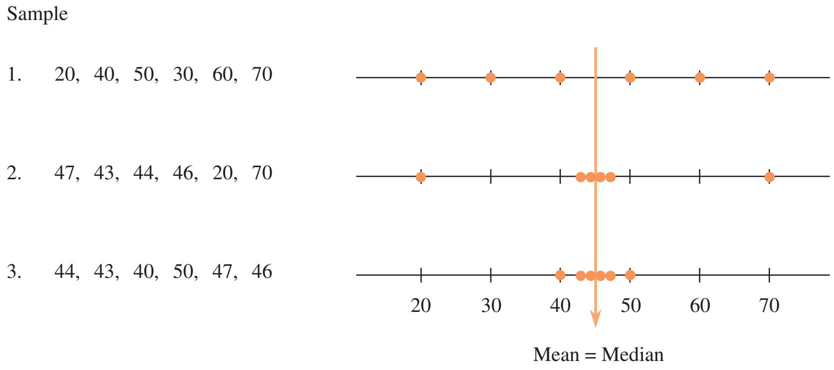


FIGURE 4.5
Three samples with the same center and different amounts of variability.

The simplest numerical measure of variability is the range.

DEFINITION

The **range** of a data set is defined as

range = largest observation − smallest observation

In general, more variability will be reflected in a larger range. However, variability is a characteristic of the entire data set, and each observation contributes to variability. The first two samples plotted in Figure 4.5 both have a range of $70 - 20 = 50$, but there is less variability in the second sample.

Deviations from the Mean

The most widely used measures of variability describe the extent to which the sample observations deviate from the sample mean $\bar{x}$. Subtracting $\bar{x}$ from each observation gives a set of deviations from the mean.

DEFINITION

The ***n*** **deviations from the sample mean** are the differences

$$(x_1 - \bar{x}), (x_2 - \bar{x}), \ldots, (x_n - \bar{x})$$

A particular deviation is positive if the corresponding x value is greater than $\bar{x}$ and negative if the x value is less than $\bar{x}$.

EXAMPLE 4.7 The Big Mac Index

● McDonald's fast-food restaurants are now found in many countries around the world. But the cost of a Big Mac varies from country to country. Table 4.2 shows data on the cost of a Big Mac (converted to U.S. dollars based on the July 2009 exchange rates) taken from the article "Cheesed Off" (*The Economist*, July 18, 2009).

TABLE 4.2 Big Mac Prices for 7 Countries

Country	Big Mac Price in U.S. Dollars
Argentina	3.02
Brazil	4.67
Chile	3.28
Colombia	3.51
Costa Rica	3.42
Peru	2.76
Uruguay	2.87

Notice that there is quite a bit of variability in the Big Mac prices.

For this data set, $\sum x = 23.53$ and $\bar{x} = \$3.36$. Table 4.3 displays the data along with the corresponding deviations, formed by subtracting $\bar{x} = 3.36$ from each observation. Three of the deviations are positive because three of the observations are larger than $\bar{x}$. The negative deviations correspond to observations that are smaller than $\bar{x}$. Some of the deviations are quite large in magnitude (1.31 and -0.60, for example), indicating observations that are far from the sample mean.

TABLE 4.3 Deviations from the Mean for the Big Mac Data

Country	Big Mac Price in U.S. Dollars	Deviations from Mean
Argentina	3.02	−0.34
Brazil	4.67	1.31
Chile	3.28	−0.08
Colombia	3.51	0.15
Costa Rica	3.42	0.06
Peru	2.76	−0.60
Uruguay	2.87	−0.49

In general, the greater the amount of variability in the sample, the larger the magnitudes (ignoring the signs) of the deviations. We now consider how to combine the deviations into a single numerical measure of variability. A first thought might be to calculate the average deviation, by adding the deviations together (this sum can be denoted compactly by $\sum(x - \bar{x})$) and then dividing by n. This does not work, though, because negative and positive deviations counteract one another in the summation.

As a result of rounding, the value of the sum of the seven deviations in Example 4.7 is $\sum(x - \bar{x}) = 0.01$. If we used even more decimal accuracy in computing $\bar{x}$ the sum would be even closer to zero.

● Data set available online

Except for the effects of rounding in computing the deviations, it is always true that

$$\sum(x - \bar{x}) = 0$$

Since this sum is zero, the average deviation is always zero and so it cannot be used as a measure of variability.

The Variance and Standard Deviation

The customary way to prevent negative and positive deviations from counteracting one another is to square them before combining. Then deviations with opposite signs but with the same magnitude, such as +2 and −2, make identical contributions to variability. The squared deviations are $(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, \ldots, (x_n - \bar{x})^2$ and their sum is

$$(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2 = \sum(x - \bar{x})^2$$

Common notation for $\sum(x - \bar{x})^2$ is S_{xx}. Dividing this sum by the sample size n gives the average squared deviation. Although this seems to be a reasonable measure of variability, we use a divisor slightly smaller than n. (The reason for this will be explained later in this section and in Chapter 9.)

DEFINITION

The **sample variance**, denoted by s^2, is the sum of squared deviations from the mean divided by $n - 1$. That is,

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

The **sample standard deviation** is the positive square root of the sample variance and is denoted by s.

A large amount of variability in the sample is indicated by a relatively large value of s^2 or s, whereas a value of s^2 or s close to zero indicates a small amount of variability. Notice that whatever unit is used for x (such as pounds or seconds), the squared deviations and therefore s^2 are in squared units. Taking the square root gives a measure expressed in the same units as x. Thus, for a sample of heights, the standard deviation might be $s = 3.2$ inches, and for a sample of textbook prices, it might be $s = \$12.43$.

EXAMPLE 4.8 Big Mac Revisited

Let's continue using the Big Mac data and the computed deviations from the mean given in Example 4.7 to calculate the sample variance and standard deviation. Table 4.4 shows the observations, deviations from the mean, and squared deviations. Combining the squared deviations to compute the values of s^2 and s gives

$$\sum(x - \bar{x}) = S_{xx} = 2.4643$$

and

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1} = \frac{2.4643}{7 - 1} = \frac{2.4643}{6} = 0.4107$$

$$s = \sqrt{0.4107} = 0.641$$

Step-by-Step technology instructions available online

TABLE 4.4 Deviations and Squared Deviations for the Big Mac Data

Big Mac Price in U.S. Dollars	Deviations from Mean	Squared Deviations
3.02	−0.34	0.1156
4.67	1.31	1.7161
3.28	−0.08	0.0064
3.51	0.15	0.0225
3.42	0.06	0.0036
2.76	−0.60	0.3600
2.87	−0.49	0.2401
		$\sum(x - \bar{x})^2 = 2.4643$

The computation of s^2 can be a bit tedious, especially if the sample size is large. Fortunately, many calculators and computer software packages compute the variance and standard deviation upon request. One commonly used statistical computer package is Minitab. The output resulting from using the Minitab Describe command with the Big Mac data follows. Minitab gives a variety of numerical descriptive measures, including the mean, the median, and the standard deviation.

```
Descriptive Statistics: Big Mac Price in U.S. Dollars
Variable       N      Mean     SE Mean   StDev   Minimum   Q1      Median
Big Mac Price  7      3.361    0.242     0.641   2.760     2.870   3.280

Variable       Q3     Maximum
Big Mac Price  3.510  4.670
```

The standard deviation can be informally interpreted as the size of a "typical" or "representative" deviation from the mean. Thus, in Example 4.8, a typical deviation from $\bar{x}$ is about 0.641; some observations are closer to $\bar{x}$ than 0.641 and others are farther away. We computed $s = 0.641$ in Example 4.8 without saying whether this value indicated a large or a small amount of variability. At this point, it is better to use s for comparative purposes than for an absolute assessment of variability. If Big Mac prices for a different group of countries resulted in a standard deviation of $s = 1.25$ (this is the standard deviation for all 45 countries for which Big Mac data was available) then we would conclude that our original sample has much less variability than the data set consisting of all 45 countries.

There are measures of variability for the entire population that are analogous to s^2 and s for a sample. These measures are called the **population variance** and the **population standard deviation** and are denoted by $\boldsymbol{\sigma^2}$ and $\boldsymbol{\sigma}$, respectively. (We again use a lowercase Greek letter for a population characteristic.)

Notation

s^2	sample variance
σ^2	population variance
s	sample standard deviation
σ	population standard deviation

In many statistical procedures, we would like to use the value of σ, but unfortunately it is not usually known. Therefore, in its place we must use a value computed from the sample that we hope is close to σ (i.e., a good *estimate* of σ). We use the divisor $(n - 1)$ in s^2 rather than n because, on average, the resulting value tends to be a bit closer to σ^2. We will say more about this in Chapter 9.

An alternative rationale for using $(n - 1)$ is based on the property $\sum(x - \bar{x}) = 0$. Suppose that $n = 5$ and that four of the deviations are

$$x_1 - \bar{x} = -4 \quad x_2 - \bar{x} = 6 \quad x_3 - \bar{x} = 1 \quad x_5 - \bar{x} = -8$$

Then, because the sum of these four deviations is -5, the remaining deviation must be $x_4 - \bar{x} = 5$ (so that the sum of all five is zero). Although there are five deviations, only four of them contain independent information about variability. More generally, once any $(n - 1)$ of the deviations are available, the value of the remaining deviation is determined. The n deviations actually contain only $(n - 1)$ independent pieces of information about variability. Statisticians express this by saying that s^2 and s are based on $(n - 1)$ *degrees of freedom* (df).

The Interquartile Range

As with $\bar{x}$, the value of s can be greatly affected by the presence of even a single unusually small or large observation. The *interquartile range* is a measure of variability that is resistant to the effects of outliers. It is based on quantities called *quartiles.* The *lower quartile* separates the bottom 25% of the data set from the upper 75%, and the *upper quartile* separates the top 25% from the bottom 75%. The *middle quartile* is the median, and it separates the bottom 50% from the top 50%. Figure 4.6 illustrates the locations of these quartiles for a smoothed histogram.

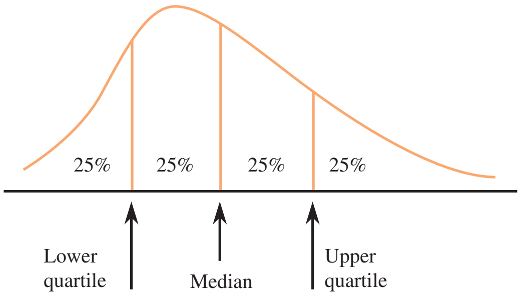


FIGURE 4.6
The quartiles for a smoothed histogram.

The quartiles for sample data are obtained by dividing the n ordered observations into a lower half and an upper half; if n is odd, the median is excluded from both halves. The two extreme quartiles are then the medians of the two halves. (Note: The median is only temporarily excluded for the purpose of computing quartiles. It is not excluded from the data set.)

DEFINITION*

lower quartile = median of the lower half of the sample
upper quartile = median of the upper half of the sample
(If n is odd, the median of the entire sample is excluded from both halves when computing quartiles.)

The **interquartile range (iqr)**, a measure of variability that is not as sensitive to the presence of outliers as the standard deviation, is given by
iqr = upper quartile − lower quartile

*There are several other sensible ways to define quartiles. Some calculators and software packages use an alternative definition.

The resistant nature of the interquartile range follows from the fact that up to 25% of the smallest sample observations and up to 25% of the largest sample observations can be made more extreme without affecting the value of the interquartile range.

EXAMPLE 4.9 Higher Education

● *The Chronicle of Higher Education* (Almanac Issue, 2009–2010) published the accompanying data on the percentage of the population with a bachelor's or higher degree in 2007 for each of the 50 U.S. states and the District of Columbia. The 51 data values are

21	27	26	19	30	35	35	26	47	26	27	30
24	29	22	24	29	20	20	27	35	38	25	31
19	24	27	27	23	34	34	25	32	26	26	24
22	28	26	30	23	25	22	25	29	33	34	30
17	25	23									

Figure 4.7 gives a stem-and-leaf display (using repeated stems) of the data. The smallest value in the data set is 17% (West Virginia), and two values stand out on the high end—38% (Massachusetts) and 47% (District of Columbia).

N = 51
Leaf Unit = 1.0

1	7
1	99
2	001
2	222333
2	444455555
2	66666677777
2	8999
3	00001
3	23
3	444555
3	
3	8
4	
4	
4	
4	7

FIGURE 4.7
Stem-and-leaf display: Percent with bachelor's or higher degree

To compute the quartiles and the interquartile range, we first order the data and use the median to divide the data into a lower half and an upper half. Because there is an odd number of observations ($n = 51$), the median is excluded from both the upper and lower halves when computing the quartiles.

Ordered Data

Lower Half:	17	19	19	20	20	21	22	22	22	23
23	23	**24**	24	24	24	25	25	25	25	25
26	26	26	26							
Median:		**26**								
Upper Half:	26	27	27	27	27	27	28	29	29	29
30	30	**30**	30	31	32	33	34	34	34	35
35	35	38	47							

Each half of the sample contains 25 observations. The lower quartile is just the median of the lower half of the sample (24 for this data set), and the upper quartile is the median of the upper half (30 for this data set). This gives

lower quartile = 24
upper quartile = 30
iqr = 30 − 24 = 6

The sample mean and standard deviation for this data set are 27.18 and 5.53, respectively. If we were to change the two largest values from 38 and 47 to 58 and 67 (so that they still remain the two largest values), the median and interquartile range would not be affected, whereas the mean and the standard deviation would change to 27.96 and 8.40, respectively. The value of the interquartile range is not affected by a few extreme values in the data set.

● Data set available online

The **population interquartile range** is the difference between the upper and lower population quartiles. If a histogram of the data set under consideration (whether a population or a sample) can be reasonably well approximated by a normal curve, then the relationship between the standard deviation (sd) and the interquartile range is roughly sd = iqr/1.35. A value of the standard deviation much larger than iqr/1.35 suggests a distribution with heavier (or longer) tails than a normal curve. For the degree data of Example 4.9, we had $s = 5.53$, whereas iqr/1.35 = 6/1.35 = 4.44. This suggests that the distribution of data values in Example 4.9 is indeed heavy-tailed compared to a normal curve. This can be seen in the stem-and-leaf display of Figure 4.7.

EXERCISES 4.17 - 4.31

4.17 ● The following data are cost (in cents) per ounce for nine different brands of sliced Swiss cheese (www.consumerreports.org):

29 62 37 41 70 82 47 52 49

a. Compute the variance and standard deviation for this data set.

b. If a very expensive cheese with a cost per slice of 150 cents was added to the data set, how would the values of the mean and standard deviation change?

4.18 ● Cost per serving (in cents) for six high-fiber cereals rated very good and for nine high-fiber cereals rated good by *Consumer Reports* are shown below. Write a few sentences describing how these two data sets differ with respect to center and variability. Use summary statistics to support your statements.

Cereals Rated Very Good

46 49 62 41 19 77

Cereals Rated Good

71 30 53 53 67 43 48 28 54

4.19 ● Combining the cost-per-serving data for high-fiber cereals rated very good and those rated good from the previous exercise gives the following data set:

46 49 62 41 19 77 71 30
53 53 67 43 48 28 54

a. Compute the quartiles and the interquartile range for this combined data set.

b. Compute the interquartile range for just the cereals rated good. Is this value greater than, less than, or about equal to the interquartile range computed in Part (a)?

4.20 ● The paper "Caffeinated Energy Drinks—A Growing Problem" (*Drug and Alcohol Dependence* [2009]: 1–10) gave the accompanying data on caffeine per ounce for eight top-selling energy drinks and for 11 high-caffeine energy drinks:

Top-Selling Energy Drinks

9.6 10.0 10.0 9.0 10.9 8.9 9.5 9.1

High-Caffeine Energy Drinks

21.0 25.0 15.0 21.5 35.7 15.0
33.3 11.9 16.3 31.3 30.0

The mean caffeine per ounce is clearly higher for the high-caffeine energy drinks, but which of the two groups of energy drinks (top-selling or high-caffeine) is the most variable with respect to caffeine per ounce? Justify your choice.

4.21 ● The Insurance Institute for Highway Safety (www.iihs.org, June 11, 2009) published data on repair costs for cars involved in different types of accidents. In one study, seven different 2009 models of mini- and micro-cars were driven at 6 mph straight into a fixed barrier. The following table gives the cost of repairing damage to the bumper for each of the seven models:

Model	Repair Cost
Smart Fortwo	$1,480
Chevrolet Aveo	$1,071
Mini Cooper	$2,291
Toyota Yaris	$1,688
Honda Fit	$1,124
Hyundai Accent	$3,476
Kia Rio	$3,701

a. Compute the values of the variance and standard deviation. The standard deviation is fairly large. What does this tell you about the repair costs?

Bold exercises answered in back ● Data set available online ✦ Video Solution available

b. The Insurance Institute for Highway Safety (referenced in the previous exercise) also gave bumper repair costs in a study of six models of minivans (December 30, 2007). Write a few sentences describing how mini- and micro-cars and minivans differ with respect to typical bumper repair cost and bumper repair cost variability.

Model	Repair Cost
Honda Odyssey	$1,538
Dodge Grand Caravan	$1,347
Toyota Sienna	$840
Chevrolet Uplander	$1,631
Kia Sedona	$1,176
Nissan Quest	$1,603

4.22 ● Consumer Reports Health (www.consumerreports.org/health) reported the accompanying caffeine concentration (mg/cup) for 12 brands of coffee:

Coffee Brand	Caffeine concentration (mg/cup)
Eight O'Clock	140
Caribou	195
Kickapoo	155
Starbucks	115
Bucks Country Coffee Co.	195
Archer Farms	180
Gloria Jean's Coffees	110
Chock Full o'Nuts	110
Peet's Coffee	130
Maxwell House	55
Folgers	60
Millstone	60

Compute the values of the quartiles and the interquartile range for this data set.

4.23 ● The accompanying data on number of minutes used for cell phone calls in 1 month was generated to be consistent with summary statistics published in a report of a marketing study of San Diego residents (TeleTruth, March 2009):

189 0 189 177 106 201 0 212 0 306
0 0 59 224 0 189 142 83 71 165
236 0 142 236 130

a. Compute the values of the quartiles and the interquartile range for this data set.

b. Explain why the lower quartile is equal to the minimum value for this data set. Will this be the case for every data set? Explain.

4.24 Give two sets of five numbers that have the same mean but different standard deviations, and give two sets of five numbers that have the same standard deviation but different means.

4.25 Going back to school can be an expensive time for parents—second only to the Christmas holiday season in terms of spending (*San Luis Obispo Tribune*, August 18, 2005). Parents spend an average of $444 on their children at the beginning of the school year stocking up on clothes, notebooks, and even iPods. Of course, not every parent spends the same amount of money and there is some variation. Do you think a data set consisting of the amount spent at the beginning of the school year for each student at a particular elementary school would have a large or a small standard deviation? Explain.

4.26 The article "Rethink Diversification to Raise Returns, Cut Risk" (*San Luis Obispo Tribune*, January 21, 2006) included the following paragraph:

> In their research, Mulvey and Reilly compared the results of two hypothetical portfolios and used actual data from 1994 to 2004 to see what returns they would achieve. The first portfolio invested in Treasury bonds, domestic stocks, international stocks, and cash. Its 10-year average annual return was 9.85% and its volatility—measured as the standard deviation of annual returns—was 9.26%. When Mulvey and Reilly shifted some assets in the portfolio to include funds that invest in real estate, commodities, and options, the 10-year return rose to 10.55% while the standard deviation fell to 7.97%. In short, the more diversified portfolio had a slightly better return and much less risk.

Explain why the standard deviation is a reasonable measure of volatility and why it is reasonable to interpret a smaller standard deviation as meaning less risk.

4.27 ● The U.S. Department of Transportation reported the accompanying data (see next page) on the number of speeding-related crash fatalities during holiday periods for the years from 1994 to 2003 (*Traffic Safety Facts*, July 20, 2005).

a. Compute the standard deviation for the New Year's Day data.

b. Without computing the standard deviation of the Memorial Day data, explain whether the standard deviation for the Memorial Day data would be larger

Data for Exercise 4.27

	Speeding-Related Fatalities									
Holiday Period	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
New Year's Day	141	142	178	72	219	138	171	134	210	70
Memorial Day	193	178	185	197	138	183	156	190	188	181
July 4th	178	219	202	179	169	176	219	64	234	184
Labor Day	183	188	166	179	162	171	180	138	202	189
Thanksgiving	212	198	218	210	205	168	187	217	210	202
Christmas	152	129	66	183	134	193	155	210	60	198

or smaller than the standard deviation of the New Year's Day data.

c. Memorial Day and Labor Day are holidays that always occur on Monday and Thanksgiving always occurs on a Thursday, whereas New Year's Day, July 4th and Christmas do not always fall on the same day of the week every year. Based on the given data, is there more or less variability in the speeding-related crash fatality numbers from year to year for same day of the week holiday periods than for holidays that can occur on different days of the week? Support your answer with appropriate measures of variability.

4.28 **The Ministry of Health and Long-Term Care in Ontario, Canada,** publishes information on the time that patients must wait for various medical procedures on its web site (www.health.gov.on.ca). For two cardiac procedures completed in fall of 2005, the following information was provided:

Procedure	Number of Completed Procedures	Median Wait Time (days)	Mean Wait Time (days)	90% Completed Within (days)
Angioplasty	847	14	18	39
Bypass surgery	539	13	19	42

a. Which of the following must be true for the lower quartile of the data set consisting of the 847 wait times for angioplasty?
   i. The lower quartile is less than 14.
   ii. The lower quartile is between 14 and 18.
   iii. The lower quartile is between 14 and 39.
   iv. The lower quartile is greater than 39.

b. Which of the following must be true for the upper quartile of the data set consisting of the 539 wait times for bypass surgery?
   i. The upper quartile is less than 13.
   ii. The upper quartile is between 13 and 19.
   iii. The upper quartile is between 13 and 42.
   iv. The upper quartile is greater than 42.

c. Which of the following must be true for the number of days for which only 5% of the bypass surgery wait times would be longer?
   i. It is less than 13.
   ii. It is between 13 and 19.
   iii. It is between 13 and 42.
   iv. It is greater than 42.

4.29 ● The accompanying table shows the low price, the high price, and the average price of homes sold in 15 communities in San Luis Obispo County between January 1, 2004, and August 1, 2004 (*San Luis Obispo Tribune*, September 5, 2004):

Community	Average Price	Number Sold	Low	High
Cayucos	$937,366	31	$380,000	$2,450,000
Pismo Beach	$804,212	71	$439,000	$2,500,000
Cambria	$728,312	85	$340,000	$2,000,000
Avila Beach	$654,918	16	$475,000	$1,375,000
Morro Bay	$606,456	114	$257,000	$2,650,000
Arroyo Grande	$595,577	214	$178,000	$1,526,000
Templeton	$578,249	89	$265,000	$2,350,000
San Luis Obispo	$557,628	277	$258,000	$2,400,000
Nipomo	$528,572	138	$263,000	$1,295,000
Los Osos	$511,866	123	$140,000	$3,500,000
Santa Margarita	$430,354	22	$290,000	$583,000
Atascadero	$420,603	270	$140,000	$1,600,000
Grover Beach	$416,405	97	$242,000	$720,000
Paso Robles	$412,584	439	$170,000	$1,575,000
Oceano	$390,354	59	$177,000	$1,350,000

a. Explain why the average price for the combined areas of Los Osos and Morro Bay is not just the average of $511,866 and $606,456.

b. Houses sold in Grover Beach and Paso Robles have very similar average prices. Based on the other information given, which is likely to have the higher standard deviation for price?

c. Consider houses sold in Grover Beach and Paso Robles. Based on the other information given, which is likely to have the higher median price?

4.30 ● In 1997, a woman sued a computer keyboard manufacturer, charging that her repetitive stress injuries were caused by the keyboard ***(Genessey v. Digital Equipment Corporation)***. The jury awarded about $3.5 million for pain and suffering, but the court then set aside that award as being unreasonable compensation. In making this determination, the court identified a "normative" group of 27 similar cases and specified a reasonable award as one within 2 standard deviations of the mean of the awards in the 27 cases. The 27 award amounts were (in thousands of dollars)

37	60	75	115	135	140	149	150
238	290	340	410	600	750	750	750
1050	1100	1139	1150	1200	1200	1250	1576
1700	1825	2000					

What is the maximum possible amount that could be awarded under the "2-standard deviations rule?"

4.31 ● The standard deviation alone does not measure relative variation. For example, a standard deviation of $1 would be considered large if it is describing the variability from store to store in the price of an ice cube tray. On the other hand, a standard deviation of $1 would be considered small if it is describing store-to-store variability in the price of a particular brand of freezer. A quantity designed to give a relative measure of variability is the *coefficient of variation*. Denoted by CV, the coefficient of variation expresses the standard deviation as a percentage of the mean. It is defined by the formula $CV = 100\left(\frac{s}{\bar{x}}\right)$. Consider two samples. Sample 1 gives the actual weight (in ounces) of the contents of cans of pet food labeled as having a net weight of 8 ounces. Sample 2 gives the actual weight (in pounds) of the contents of bags of dry pet food labeled as having a net weight of 50 pounds. The weights for the two samples are

Sample 1	8.3	7.1	7.6	8.1	7.6
	8.3	8.2	7.7	7.7	7.5
Sample 2	52.3	50.6	52.1	48.4	48.8
	47.0	50.4	50.3	48.7	48.2

a. For each of the given samples, calculate the mean and the standard deviation.

b. Compute the coefficient of variation for each sample. Do the results surprise you? Why or why not?

Bold exercises answered in back ● Data set available online ✦ Video Solution available

4.3 Summarizing a Data Set: Boxplots

In Sections 4.1 and 4.2, we looked at ways of describing the center and variability of a data set using numerical measures. It would be nice to have a method of summarizing data that gives more detail than just a measure of center and spread and yet less detail than a stem-and-leaf display or histogram. A *boxplot* is one way to do this. A boxplot is compact, yet it provides information about the center, spread, and symmetry or skewness of the data. We will consider two types of boxplots: the skeletal boxplot and the modified boxplot.

Construction of a Skeletal Boxplot

1. Draw a horizontal (or vertical) measurement scale.
2. Construct a rectangular box with a left (or lower) edge at the lower quartile and a right (or upper) edge at the upper quartile. The box width is then equal to the iqr.
3. Draw a vertical (or horizontal) line segment inside the box at the location of the median.
4. Extend horizontal (or vertical) line segments, called whiskers, from each end of the box to the smallest and largest observations in the data set.

EXAMPLE 4.10 Revisiting the Degree Data

Let's reconsider the data on percentage of the population with a bachelor's or higher degree for the 50 U.S. states and the District of Columbia (Example 4.9). The ordered observations are

Ordered Data

Lower Half:	17	19	19	20	20	21	22	22	22	23
23	23	**24**	24	24	24	25	25	25	25	25
26	26	26	26							
Median:		**26**								
Upper Half:	26	27	27	27	27	27	28	29	29	29
30	30	**30**	30	31	32	33	34	34	34	35
35	35	38	47							

To construct a boxplot of these data, we need the following information: the smallest observation, the lower quartile, the median, the upper quartile, and the largest observation. This collection of summary measures is often referred to as a **five-number summary.** For this data set we have

smallest observation = 17
lower quartile = median of the lower half = 24
median = 26th observation in the ordered list = 26
upper quartile = median of the upper half = 30
largest observation = 47

Figure 4.8 shows the corresponding boxplot. The median line is somewhat closer to the lower edge of the box than to the upper edge, suggesting a concentration of values in the lower part of the middle half. The upper whisker is longer than the lower whisker. These observations are consistent with the stem-and-leaf display of Figure 4.7.

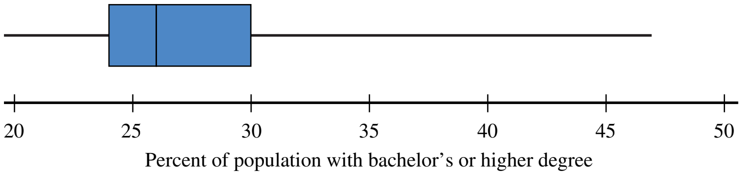


FIGURE 4.8
Skeletal boxplot for the degree data of Example 4.10.

The sequence of steps used to construct a skeletal boxplot is easily modified to give information about outliers.

DEFINITION

An observation is an **outlier** if it is more than 1.5(iqr) away from the nearest quartile (the nearest end of the box).

An outlier is **extreme** if it is more than 3(iqr) from the nearest quartile and it is **mild** otherwise.

A **modified boxplot** represents mild outliers by solid circles and extreme outliers by open circles, and the whiskers extend on each end to the most extreme observations that are *not* outliers.

Construction of a Modified Boxplot

1. Draw a horizontal (or vertical) measurement scale.
2. Construct a rectangular box with a left (or lower) edge at the lower quartile and right (or upper) edge at the upper quartile. The box width is then equal to the iqr.
3. Draw a vertical (or horizontal) line segment inside the box at the location of the median.
4. Determine if there are any mild or extreme outliers in the data set.
5. Draw whiskers that extend from each end of the box to the most extreme observation that is *not* an outlier.
6. Draw a solid circle to mark the location of any mild outliers in the data set.
7. Draw an open circle to mark the location of any extreme outliers in the data set.

EXAMPLE 4.11 Golden Rectangles

● The accompanying data came from an anthropological study of rectangular shapes (*Lowie's Selected Papers in Anthropology,* Cora Dubios, ed. [Berkeley, CA: University of California Press, 1960]: 137–142). Observations were made on the variable x = width/length for a sample of $n = 20$ beaded rectangles used in Shoshoni Indian leather handicrafts:

.553	.570	.576	.601	.606	.606	.609	.611	.615	.628
.654	.662	.668	.670	.672	.690	.693	.749	.844	.933

The quantities needed for constructing the modified boxplot follow:

median = .641 iqr = .681 − .606 = .075
lower quartile = .606 1.5(iqr) = .1125
upper quartile = .681 3(iqr) = .225

Thus,

$$\text{(upper quartile)} + 1.5(\text{iqr}) = .681 + .1125 = .7935$$
$$\text{(lower quartile)} - 1.5(\text{iqr}) = .606 - .1125 = .4935$$

So 0.844 and 0.933 are both outliers on the upper end (because they are larger than 0.7935), and there are no outliers on the lower end (because no observations are smaller than 0.4935). Because

$$\text{(upper quartile)} + 3(\text{iqr}) = 0.681 + 0.225 = 0.906$$

0.933 is an extreme outlier and 0.844 is only a mild outlier. The upper whisker extends to the largest observation that is not an outlier, 0.749, and the lower whisker extends to 0.553. The boxplot is shown in Figure 4.9. The median line is not at the center of the box, so there is a slight asymmetry in the middle half of the data. However, the most striking feature is the presence of the two outliers. These two x values considerably exceed the "golden ratio" of 0.618, used since antiquity as an aesthetic standard for rectangles.

Step-by-Step technology instructions available online

● Data set available online

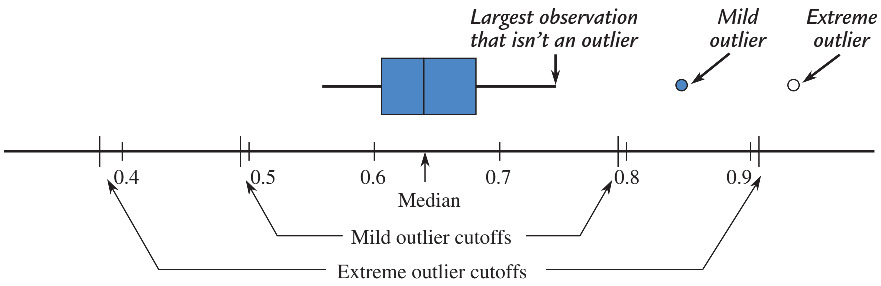


FIGURE 4.9
Boxplot for the rectangle data in Example 4.11.

EXAMPLE 4.12 Another Look at Big Mac Prices

Big Mac prices in U.S. dollars for 45 different countries were given in the article "Cheesed Off" first introduced in Example 4.7. The 45 Big Mac prices were:

3.57	3.01	3.97	4.67	3.80	3.64	3.28	1.83	3.51	3.42	3.92
5.89	3.04	2.36	4.92	1.72	3.89	5.20	2.21	3.98	3.54	3.24
3.06	1.99	2.48	3.54	7.03	2.28	2.76	2.09	2.66	2.31	2.93
3.03	2.37	2.91	1.83	5.57	6.39	2.31	1.93	3.80	2.72	1.70
2.87										

Figure 4.10 shows a Minitab boxplot for the Big Mac price data. Note that the upper whisker is longer than the lower whisker and that there are two outliers on the high end (Norway with a Big Mac price of $7.04 and Switzerland with a price of $6.29).

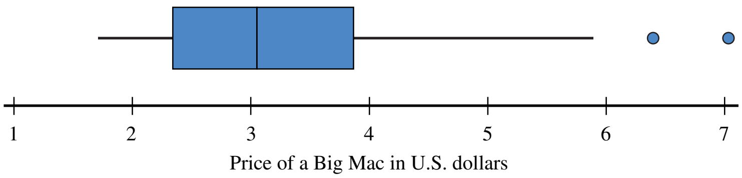


FIGURE 4.10
Minitab boxplot of the Big Mac price data of Example 4.12.

Note that Minitab does not distinguish between mild outliers and extreme outliers in the boxplot. For the Big Mac price data,

$$\text{lower quartile} = 2.335$$
$$\text{upper quartile} = 3.845$$
$$\text{iqr} = 3.845 - 2.335 = 1.510$$

Then

$$1.5(\text{iqr}) = 2.265$$
$$3(\text{iqr}) = 4.530$$

We can compute outlier boundaries as follows:

$$\text{upper quartile} + 1.5(\text{iqr}) = 3.845 + 2.265 = 6.110$$
$$\text{upper quartile} + 3(\text{iqr}) = 3.845 + 4.530 = 8.375$$

The observation for Switzerland (6.39) is a mild outlier because it is greater than 6.110 (the upper quartile + 1.5(iqr)) but less than 8.375 (the upper quartile + 3(iqr)). The observation for Norway is also a mild outlier. There are no extreme outliers in this data set.

With two or more data sets consisting of observations on the same variable (for example, fuel efficiencies for four types of car or weight gains for a control group and a treatment group), **comparative boxplots** (more than one boxplot drawn using the same scale) can tell us a lot about similarities and differences between the data sets.

EXAMPLE 4.13 NBA Salaries Revisited

The 2009–2010 salaries of NBA players published on the web site hoopshype.com were used to construct the comparative boxplot of the salary data for five teams shown in Figure 4.11.

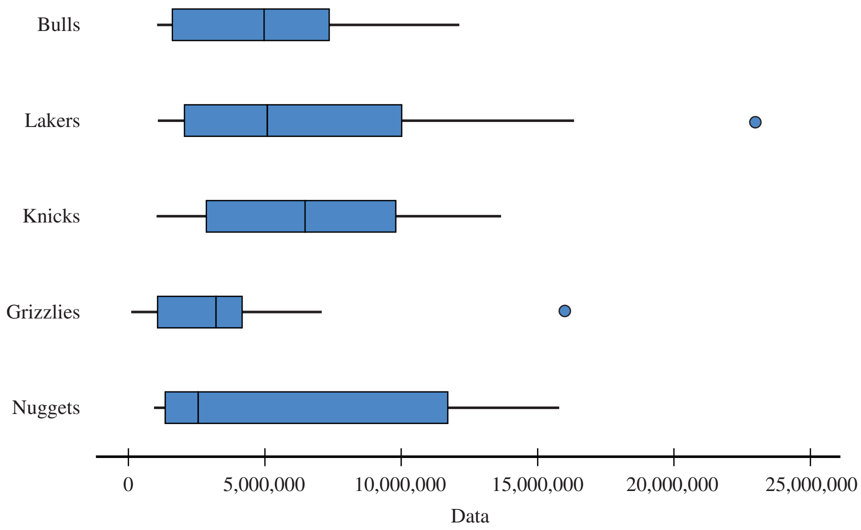


FIGURE 4.11
Comparative boxplot for salaries for five NBA teams.

The comparative boxplot reveals some interesting similarities and differences in the salary distributions of the five teams. The minimum salary is lower for the Grizzlies, but is about the same for the other four teams. The median salary was lowest for the Nuggets—in fact the median for the Nuggets is about the same as the lower quartile for the Knicks and the Lakers, indicating that half of the players on the Nuggets have salaries less than about $2.5 million, whereas only about 25% of the Knicks and the Lakers have salaries less than about $2.5 million. The Lakers had the player with by far the highest salary. The Grizzlies and the Lakers were the only teams that had any salary outliers. With the exception of one highly paid player, salaries for players on the Grizzlies team were noticeably lower than for the other four teams.

EXERCISES 4.32 - 4.37

4.32 Based on a large national sample of working adults, the U.S. Census Bureau reports the following information on travel time to work for those who do not work at home:

lower quartile = 7 minutes
median = 18 minutes
upper quartile = 31 minutes

Also given was the mean travel time, which was reported as 22.4 minutes.

a. Is the travel time distribution more likely to be approximately symmetric, positively skewed, or negatively skewed? Explain your reasoning based on the given summary quantities.

b. Suppose that the minimum travel time was 1 minute and that the maximum travel time in the sample was 205 minutes. Construct a skeletal boxplot for the travel time data.

c. Were there any mild or extreme outliers in the data set? How can you tell?

4.33 ● The report "Who Moves? Who Stays Put? Where's Home?" (*Pew Social and Demographic Trends*, December 17, 2008) gave the accompanying data for the 50 U.S. states on the percentage of the population that was born in the state and is still living there. The data values have been arranged in order from largest to smallest.

75.8	71.4	69.6	69.0	68.6	67.5	66.7	66.3	66.1	66.0	66.0
65.1	64.4	64.3	63.8	63.7	62.8	62.6	61.9	61.9	61.5	61.1
59.2	59.0	58.7	57.3	57.1	55.6	55.6	55.5	55.3	54.9	54.7
54.5	54.0	54.0	53.9	53.5	52.8	52.5	50.2	50.2	48.9	48.7
48.6	47.1	43.4	40.4	35.7	28.2					

a. Find the values of the median, the lower quartile, and the upper quartile.

b. The two smallest values in the data set are 28.2 (Alaska) and 35.7 (Wyoming). Are these two states outliers?

c. Construct a boxplot for this data set and comment on the interesting features of the plot.

4.34 ● The National Climate Data Center gave the accompanying annual rainfall (in inches) for Medford, Oregon, from 1950 to 2008 (www.ncdc.noaa.gov/oa/climate/research/cag3/city.html):

28.84	20.15	18.88	25.72	16.42	20.18	28.96	20.72	23.58
10.62	20.85	19.86	23.34	19.08	29.23	18.32	21.27	18.93
15.47	20.68	23.43	19.55	20.82	19.04	18.77	19.63	12.39
22.39	15.95	20.46	16.05	22.08	19.44	30.38	18.79	10.89
17.25	14.95	13.86	15.30	13.71	14.68	15.16	16.77	12.33
21.93	31.57	18.13	28.87	16.69	18.81	15.15	18.16	19.99
19.00	23.97	21.99	17.25	14.07				

a. Compute the quartiles and the interquartile range.

b. Are there outliers in this data set? If so, which observations are mild outliers? Which are extreme outliers?

c. Draw a boxplot for this data set that shows outliers.

4.35 ● The accompanying data on annual maximum wind speed (in meters per second) in Hong Kong for each year in a 45-year period were given in an article that appeared in the journal *Renewable Energy* (March 2007). Use the annual maximum wind speed data to construct a boxplot. Is the boxplot approximately symmetric?

30.3	39.0	33.9	38.6	44.6	31.4	26.7	51.9	31.9
27.2	52.9	45.8	63.3	36.0	64.0	31.4	42.2	41.1
37.0	34.4	35.5	62.2	30.3	40.0	36.0	39.4	34.4
28.3	39.1	55.0	35.0	28.8	25.7	62.7	32.4	31.9
37.5	31.5	32.0	35.5	37.5	41.0	37.5	48.6	28.1

4.36 ● Fiber content (in grams per serving) and sugar content (in grams per serving) for 18 high fiber cereals (www.consumerreports.com) are shown below.

Fiber Content

7	10	10	7	8	7	12	12	8
13	10	8	12	7	14	7	8	8

Sugar Content

11	6	14	13	0	18	9	10	19
6	10	17	10	10	0	9	5	11

a. Find the median, quartiles, and interquartile range for the fiber content data set.

b. Find the median, quartiles, and interquartile range for the sugar content data set.

c. Are there any outliers in the sugar content data set?

d. Explain why the minimum value for the fiber content data set and the lower quartile for the fiber content data set are equal.

e. Construct a comparative boxplot and use it to comment on the differences and similarities in the fiber and sugar distributions.

4.37 ● Shown here are the number of auto accidents per year for every 1000 people in each of 40 occupations (*Knight Ridder Tribune*, June 19, 2004):

Occupation	Accidents per 1000	Occupation	Accidents per 1000
Student	152	Social worker	98
Physician	109	Manual laborer	96
Lawyer	106	Analyst	95
Architect	105	Engineer	94
Real estate broker	102	Consultant	94
Enlisted military	99	Sales	93

(continued)

Occupation	Accidents per 1000	Occupation	Accidents per 1000
Military officer	91	Pharmacist	85
Nurse	90	Proprietor	84
School administrator	90	Teacher, professor	84
Skilled laborer	90	Accountant	84
Librarian	90	Law enforcement	79
Creative arts	90	Physical therapist	78
Executive	89	Veterinarian	78
Insurance agent	89	Clerical, secretary	77
Banking, finance	89	Clergy	76
Customer service	88	Homemaker	76
Manager	88	Politician	76
Medical support	87	Pilot	75
Computer-related	87	Firefighter	67
Dentist	86	Farmer	43

a. Would you recommend using the standard deviation or the iqr as a measure of variability for this data set?

b. Are there outliers in this data set? If so, which observations are mild outliers? Which are extreme outliers?

c. Draw a modified boxplot for this data set.

d. If you were asked by an insurance company to decide which, if any, occupations should be offered a professional discount on auto insurance, which occupations would you recommend? Explain.

Bold exercises answered in back ● Data set available online ✦ Video Solution available

4.4 Interpreting Center and Variability: Chebyshev's Rule, the Empirical Rule, and z Scores

The mean and standard deviation can be combined to make informative statements about how the values in a data set are distributed and about the relative position of a particular value in a data set. To do this, it is useful to be able to describe how far away a particular observation is from the mean in terms of the standard deviation. For example, we might say that an observation is 2 standard deviations above the mean or that an observation is 1.3 standard deviations below the mean.

EXAMPLE 4.14 Standardized Test Scores

Consider a data set of scores on a standardized test with a mean and standard deviation of 100 and 15, respectively. We can make the following statements:

1. Because 100 − 15 = 85, we say that a score of 85 is "1 standard deviation *below* the mean." Similarly, 100 + 15 = 115 is "1 standard deviation *above* the mean" (see Figure 4.12).

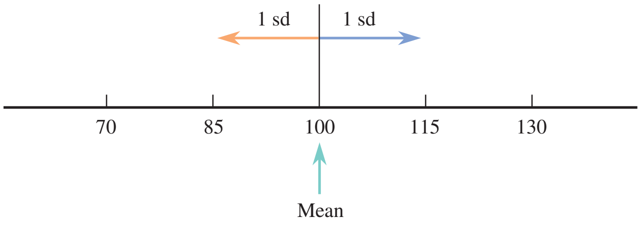


FIGURE 4.12
Values within 1 standard deviation of the mean (Example 4.14).

2. Because 2 times the standard deviation is $2(15) = 30$, and $100 + 30 = 130$ and $100 - 30 = 70$, scores between 70 and 130 are those *within* 2 standard deviations of the mean (see Figure 4.13).
3. Because $100 + (3)(15) = 145$, scores above 145 are greater than the mean by more than 3 standard deviations.

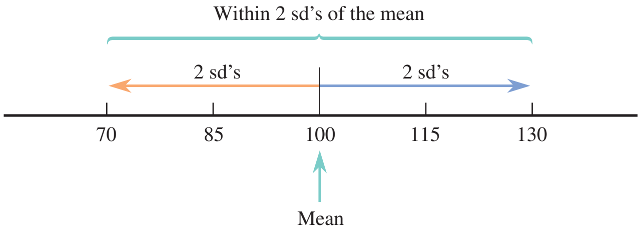


FIGURE 4.13
Values within 2 standard deviations of the mean (Example 4.14).

Sometimes in published articles, the mean and standard deviation are reported, but a graphical display of the data is not given. However, using a result called Chebyshev's Rule, it is possible to get a sense of the distribution of data values based on our knowledge of only the mean and standard deviation.

Chebyshev's Rule

Consider any number k, where $k \geq 1$. Then the percentage of observations that are within k standard deviations of the mean is at least $100\left(1 - \frac{1}{k^2}\right)\%$. Substituting selected values of k gives the following results.

Number of Standard Deviations, k	$1 - \frac{1}{k^2}$	Percentage Within k Standard Deviations of the Mean
2	$1 - \frac{1}{4} = .75$	at least 75%
3	$1 - \frac{1}{9} = .89$	at least 89%
4	$1 - \frac{1}{16} = .94$	at least 94%
4.472	$1 - \frac{1}{20} = .95$	at least 95%
5	$1 - \frac{1}{25} = .96$	at least 96%
10	$1 - \frac{1}{100} = .99$	at least 99%

EXAMPLE 4.15 Child Care for Preschool Kids

The article "Piecing Together Child Care with Multiple Arrangements: Crazy Quilt or Preferred Pattern for Employed Parents of Preschool Children?" (*Journal of Marriage and the Family* [1994]: 669–680) examined various modes of care for preschool children. For a sample of families with one preschool child, it was reported that the mean and standard deviation of child care time per week were approximately 36 hours and 12 hours, respectively. Figure 4.14 displays values that are 1, 2, and 3 standard deviations from the mean.

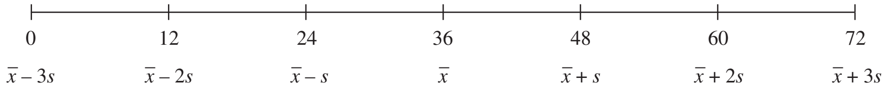


FIGURE 4.14
Measurement scale for child care time (Example 4.15).

Ariel Skelley/Blend Images/Jupiter Images

Chebyshev's Rule allows us to assert the following:

1. At least 75% of the sample observations must be between 12 and 60 hours (within 2 standard deviations of the mean).
2. Because at least 89% of the observations must be between 0 and 72, at most 11% are outside this interval. Time cannot be negative, so we conclude that at most 11% of the observations exceed 72.
3. The values 18 and 54 are 1.5 standard deviations to either side of $\bar{x}$, so using $k = 1.5$ in Chebyshev's Rule implies that at least 55.6% of the observations must be between these two values. Thus, at most 44.4% of the observations are less than 18—*not* at most 22.2%, because the distribution of values may not be symmetric.

Because Chebyshev's Rule is applicable to any data set (distribution), whether symmetric or skewed, we must be careful when making statements about the proportion above a particular value, below a particular value, or inside or outside an interval that is not centered at the mean. The rule must be used in a conservative fashion. There is another aspect of this conservatism. The rule states that *at least* 75% of the observations are within 2 standard deviations of the mean, but for many data sets substantially more than 75% of the values satisfy this condition. The same sort of understatement is frequently encountered for other values of k (numbers of standard deviations).

EXAMPLE 4.16 IQ Scores

Figure 4.15 gives a stem-and-leaf display of IQ scores of 112 children in one of the early studies that used the Stanford revision of the Binet–Simon intelligence scale (*The Intelligence of School Children,* L. M. Terman [Boston: Houghton-Mifflin, 1919]).

Summary quantities include

$$\bar{x} = 104.5 \quad s = 16.3 \quad 2s = 32.6 \quad 3s = 48.9$$

6	1
7	25679
8	0000124555668
9	0000112333446666778889
10	00011222223335666777788899999
11	00001122333344444477899
12	01111123445669
13	006
14	26
15	2

Stem: Tens
Leaf: Ones

FIGURE 4.15
Stem-and-leaf display of IQ scores used in Example 4.16.

In Figure 4.15, all observations that are within two standard deviations of the mean are shown in blue. Table 4.5 shows how Chebyshev's Rule can sometimes considerably understate actual percentages.

TABLE 4.5 Summarizing the Distribution of IQ Scores

k = Number of sd's	$\bar{x} \pm ks$	Chebyshev	Actual
2	71.9 to 137.1	at least 75%	96% (108)
2.5	63.7 to 145.3	at least 84%	97% (109)
3	55.6 to 153.4	at least 89%	100% (112)

← *the blue leaves in Figure 4.15*

Empirical Rule

The fact that statements based on Chebyshev's Rule are frequently conservative suggests that we should look for rules that are less conservative and more precise. One useful rule is the **Empirical Rule**, which can be applied whenever the distribution of data values can be reasonably well described by a normal curve (distributions that are "mound" shaped).

The Empirical Rule

If the histogram of values in a data set can be reasonably well approximated by a normal curve, then

Approximately 68% of the observations are within 1 standard deviation of the mean.

Approximately 95% of the observations are within 2 standard deviations of the mean.

Approximately 99.7% of the observations are within 3 standard deviations of the mean.

The Empirical Rule makes "approximately" instead of "at least" statements, and the percentages for k = 1, 2, and 3 standard deviations are much higher than those of Chebyshev's Rule. Figure 4.16 illustrates the percentages given by the Empirical Rule. In contrast to Chebyshev's Rule, dividing the percentages in half is permissible, because a normal curve is symmetric.

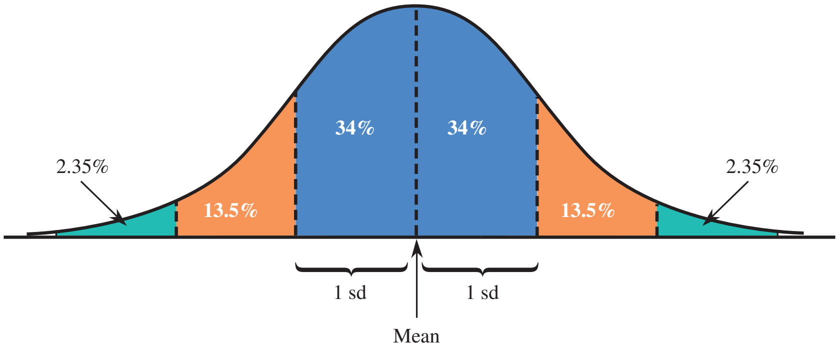


FIGURE 4.16
Approximate percentages implied by the Empirical Rule.

EXAMPLE 4.17 Heights of Mothers and the Empirical Rule

The Image Bank/Paul Thomas/ Getty Images

One of the earliest articles to argue for the wide applicability of the normal distribution was "On the Laws of Inheritance in Man. I. Inheritance of Physical Characters" (*Biometrika* [1903]: 375–462). Among the data sets discussed in the article was one consisting of 1052 measurements of the heights of mothers. The mean and standard deviation were

$$\bar{x} = 62.484 \text{ in.} \qquad s = 2.390 \text{ in.}$$

The data distribution was described as approximately normal. Table 4.6 contrasts actual percentages with those obtained from Chebyshev's Rule and the Empirical Rule.

TABLE 4.6 Summarizing the Distribution of Mothers' Heights

Number of sd's	Interval	Actual	Empirical Rule	Chebyshev Rule
1	60.094 to 64.874	72.1%	Approximately 68%	At least 0%
2	57.704 to 67.264	96.2%	Approximately 95%	At least 75%
3	55.314 to 69.654	99.2%	Approximately 99.7%	At least 89%

Clearly, the Empirical Rule is much more successful and informative in this case than Chebyshev's Rule.

Our detailed study of the normal distribution and areas under normal curves in Chapter 7 will enable us to make statements analogous to those of the Empirical Rule for values other than $k = 1$, 2, or 3 standard deviations. For now, note that it is unusual to see an observation from a normally distributed population that is farther than 2 standard deviations from the mean (only 5%), and it is very surprising to see one that is more than 3 standard deviations away. If you encountered a mother whose height was 72 inches, you might reasonably conclude that she was not part of the population described by the data set in Example 4.17.

Measures of Relative Standing

When you obtain your score after taking a test, you probably want to know how it compares to the scores of others who have taken the test. Is your score above or below the mean, and by how much? Does your score place you among the top 5% of those who took the test or only among the top 25%? Questions of this sort are answered by finding ways to measure the position of a particular value in a data set relative to all values in the set. One measure of relative standing is a *z score.*

DEFINITION

The **z score** corresponding to a particular value is

$$z \text{ score} = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

The z score tells us how many standard deviations the value is from the mean. It is positive or negative according to whether the value lies above or below the mean.

The process of subtracting the mean and then dividing by the standard deviation is sometimes referred to as *standardization,* and a z score is one example of what is called a *standardized score.*

EXAMPLE 4.18 Relatively Speaking, Which Is the Better Offer?

Suppose that two graduating seniors, one a marketing major and one an accounting major, are comparing job offers. The accounting major has an offer for \$45,000 per year, and the marketing student has an offer for \$43,000 per year. Summary information about the distribution of offers follows:

Accounting:	mean = 46,000	standard deviation = 1500
Marketing:	mean = 42,500	standard deviation = 1000

Then,

$$\text{accounting } z \text{ score} = \frac{45{,}000 - 46{,}000}{1500} = -.67$$

(so \$45,000 is .67 standard deviation below the mean), whereas

$$\text{marketing } z \text{ score} = \frac{43{,}000 - 42{,}500}{1000} = .5$$

Relative to the appropriate data sets, the marketing offer is actually more attractive than the accounting offer (although this may not offer much solace to the marketing major).

The z score is particularly useful when the distribution of observations is approximately normal. In this case, from the Empirical Rule, a z score outside the interval from -2 to $+2$ occurs in about 5% of all cases, whereas a z score outside the interval from -3 to $+3$ occurs only about 0.3% of the time.

Percentiles

A particular observation can be located even more precisely by giving the percentage of the data that fall at or below that observation. If, for example, 95% of all test scores are at or below 650, whereas only 5% are above 650, then 650 is called the *95th percentile* of the data set (or of the distribution of scores). Similarly, if 10% of all scores are at or below 400 and 90% are above 400, then the value 400 is the 10th percentile.

DEFINITION

For any particular number r between 0 and 100, the **rth percentile** is a value such that r percent of the observations in the data set fall at or below that value.

Figure 4.17 illustrates the 90th percentile. We have already met several percentiles in disguise. The median is the 50th percentile, and the lower and upper quartiles are the 25th and 75th percentiles, respectively.

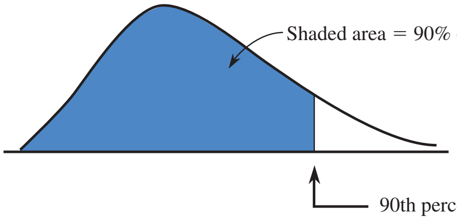


FIGURE 4.17
Ninetieth percentile for a smoothed histogram.

EXAMPLE 4.19 Head Circumference at Birth

In addition to weight and length, head circumference is another measure of health in newborn babies. The National Center for Health Statistics reports the following summary values for head circumference (in cm) at birth for boys (approximate values read from graphs on the Center for Disease Control web site):

Percentile	5	10	25	50	75	90	95
Head Circumference (cm)	32.2	33.2	34.5	35.8	37.0	38.2	38.6

Interpreting these percentiles, we know that half of newborn boys have head circumferences of less than 35.8 cm, because 35.8 is the 50th percentile (the median). The middle 50% of newborn boys have head circumferences between 34.5 cm and 37.0 cm, with about 25% of the head circumferences less than 34.5 cm and about 25% greater than 37.0 cm. We can tell that the head circumference distribution for newborn boys is not symmetric, because the 5th percentile is 3.6 cm below the median, whereas the 95th percentile is only 2.8 cm above the median. This suggests that the bottom part of the distribution stretches out more than the top part of the distribution. This would be consistent with a distribution that is negatively skewed, as shown in Figure 4.18.

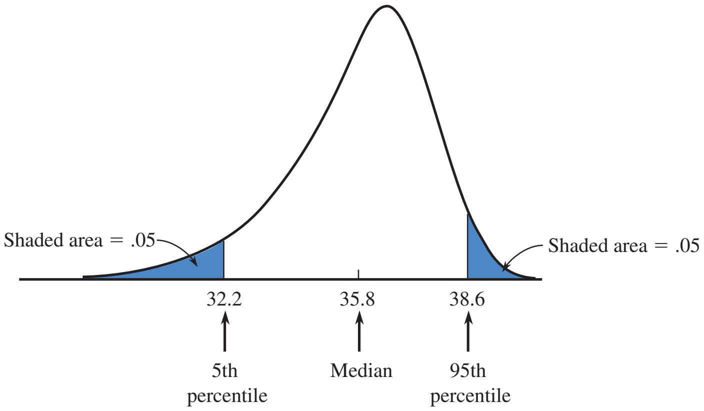


FIGURE 4.18
Negatively skewed distribution.

EXERCISES 4.38 - 4.52

4.38 The average playing time of compact discs in a large collection is 35 minutes, and the standard deviation is 5 minutes.

a. What value is 1 standard deviation above the mean? 1 standard deviation below the mean? What values are 2 standard deviations away from the mean?

b. Without assuming anything about the distribution of times, at least what percentage of the times is between 25 and 45 minutes?

c. Without assuming anything about the distribution of times, what can be said about the percentage of times that are either less than 20 minutes or greater than 50 minutes?

d. Assuming that the distribution of times is approximately normal, about what percentage of times are between 25 and 45 minutes? less than 20 minutes or greater than 50 minutes? less than 20 minutes?

4.39 ✦ In a study investigating the effect of car speed on accident severity, 5000 reports of fatal automobile accidents were examined, and the vehicle speed at impact was recorded for each one. For these 5000 accidents, the average speed was 42 mph and the standard deviation was 15 mph. A histogram revealed that the vehicle speed at impact distribution was approximately normal.

a. Roughly what proportion of vehicle speeds were between 27 and 57 mph?

b. Roughly what proportion of vehicle speeds exceeded 57 mph?

4.40 The U.S. Census Bureau (2000 census) reported the following relative frequency distribution for travel time to work for a large sample of adults who did not work at home:

Travel Time (minutes)	Relative Frequency
0 to <5	.04
5 to <10	.13
10 to <15	.16
15 to <20	.17
20 to <25	.14
25 to <30	.05
30 to <35	.12
35 to <40	.03
40 to <45	.03
45 to <60	.06
60 to <90	.05
90 or more	.02

a. Draw the histogram for the travel time distribution. In constructing the histogram, assume that the last interval in the relative frequency distribution (90 or more) ends at 200; so the last interval is 90 to <200. Be sure to use the density scale to determine the heights of the bars in the histogram because not all the intervals have the same width.

b. Describe the interesting features of the histogram from Part (a), including center, shape, and spread.

c. Based on the histogram from Part (a), would it be appropriate to use the Empirical Rule to make statements about the travel time distribution? Explain why or why not.

d. The approximate mean and standard deviation for the travel time distribution are 27 minutes and 24 minutes, respectively. Based on this mean and standard deviation and the fact that travel time cannot be negative, explain why the travel time distribution could not be well approximated by a normal curve.

e. Use the mean and standard deviation given in Part (d) and Chebyshev's Rule to make a statement about

i. the percentage of travel times that were between 0 and 75 minutes

ii. the percentage of travel times that were between 0 and 47 minutes

f. How well do the statements in Part (e) based on Chebyshev's Rule agree with the actual percentages for the travel time distribution? (Hint: You can estimate the actual percentages from the given relative frequency distribution.)

4.41 Mobile homes are tightly constructed for energy conservation. This can lead to a buildup of indoor pollutants. The paper "A Survey of Nitrogen Dioxide Levels Inside Mobile Homes" (*Journal of the Air Pollution Control Association* [1988]: 647–651) discussed various aspects of NO_2 concentration in these structures.

a. In one sample of mobile homes in the Los Angeles area, the mean NO_2 concentration in kitchens during the summer was 36.92 ppb, and the standard deviation was 11.34. Making no assumptions about the shape of the NO_2 distribution, what can be said about the percentage of observations between 14.24 and 59.60?

b. Inside what interval is it guaranteed that at least 89% of the concentration observations will lie?

c. In a sample of non–Los Angeles mobile homes, the average kitchen NO_2 concentration during the winter was 24.76 ppb, and the standard deviation was 17.20. Do these values suggest that the histogram of sample observations did not closely resemble a normal curve? (Hint: What is $\bar{x} - 2s$?)

4.42 The article "Taxable Wealth and Alcoholic Beverage Consumption in the United States" (*Psychological Reports* [1994]: 813–814) reported that the mean annual adult consumption of wine was 3.15 gallons and that the standard deviation was 6.09 gallons. Would you use the Empirical Rule to approximate the proportion of adults who consume more than 9.24 gallons (i.e., the proportion of adults whose consumption value exceeds the mean by more than 1 standard deviation)? Explain your reasoning.

4.43 A student took two national aptitude tests. The national average and standard deviation were 475 and 100, respectively, for the first test and 30 and 8, respectively, for the second test. The student scored 625 on the first test and 45 on the second test. Use z scores to determine on which exam the student performed better relative to the other test takers.

4.44 Suppose that your younger sister is applying for entrance to college and has taken the SATs. She scored at the 83rd percentile on the verbal section of the test and at the 94th percentile on the math section of the test. Because you have been studying statistics, she asks you for an interpretation of these values. What would you tell her?

4.45 A sample of concrete specimens of a certain type is selected, and the compressive strength of each specimen is determined. The mean and standard deviation are calculated as $\bar{x} = 3000$ and $s = 500$, and the sample histogram is found to be well approximated by a normal curve.

a. Approximately what percentage of the sample observations are between 2500 and 3500?

b. Approximately what percentage of sample observations are outside the interval from 2000 to 4000?

c. What can be said about the approximate percentage of observations between 2000 and 2500?

d. Why would you not use Chebyshev's Rule to answer the questions posed in Parts (a)–(c)?

4.46 The paper "Modeling and Measurements of Bus Service Reliability" (*Transportation Research* [1978]: 253–256) studied various aspects of bus service and presented data on travel times (in minutes) from several different routes. The accompanying frequency distribution is for bus travel times from origin to destination on one particular route in Chicago during peak morning traffic periods:

Travel Time	Frequency	Relative Frequency
15 to <16	4	.02
16 to <17	0	.00
17 to <18	26	.13
18 to <19	99	.49
19 to <20	36	.18
20 to <21	8	.04
21 to <22	12	.06
22 to <23	0	.00
23 to <24	0	.00
24 to <25	0	.00
25 to <26	16	.08

a. Construct the corresponding histogram.

b. Compute (approximately) the following percentiles:

i. 86th
ii. 15th
iii. 90th
iv. 95th
v. 10th

4.47 An advertisement for the "30 inch Wonder" that appeared in the September 1983 issue of the journal *Packaging* claimed that the 30 inch Wonder weighs cases and bags up to 110 pounds and provides accuracy to within 0.25 ounce. Suppose that a 50 ounce weight was repeatedly weighed on this scale and the weight readings recorded. The mean value was 49.5 ounces, and the standard deviation was 0.1. What can be said about the proportion of the time that the scale actually showed a weight that was within 0.25 ounce of the true value of 50 ounces? (Hint: Use Chebyshev's Rule.)

4.48 Suppose that your statistics professor returned your first midterm exam with only a z score written on it. She also told you that a histogram of the scores was approximately normal. How would you interpret each of the following z scores?

a. 2.2
b. 0.4
c. 1.8
d. 1.0
e. 0

4.49 The paper "Answer Changing on Multiple-Choice Tests" (*Journal of Experimental Education* [1980]: 18–21) reported that for a group of 162 college students, the average number of responses changed from the correct answer to an incorrect answer on a test containing 80 multiple-choice items was 1.4. The corresponding standard deviation was reported to be 1.5. Based on this mean and standard deviation, what can you tell about the shape of the distribution of the variable *number of answers changed from right to wrong?* What can you say about the number of students who changed at least six answers from correct to incorrect?

4.50 The average reading speed of students completing a speed-reading course is 450 words per minute (wpm). If the standard deviation is 70 wpm, find the *z* score associated with each of the following reading speeds.

a. 320 wpm c. 420 wpm
b. 475 wpm d. 610 wpm

4.51 ● The following data values are 2009 per capita expenditures on public libraries for each of the 50 U.S. states (from www.statemaster.com):

16.84	16.17	11.74	11.11	8.65	7.69	7.48
7.03	6.20	6.20	5.95	5.72	5.61	5.47
5.43	5.33	4.84	4.63	4.59	4.58	3.92
3.81	3.75	3.74	3.67	3.40	3.35	3.29
3.18	3.16	2.91	2.78	2.61	2.58	2.45
2.30	2.19	2.06	1.78	1.54	1.31	1.26
1.20	1.19	1.09	0.70	0.66	0.54	0.49
0.30	0.01					

a. Summarize this data set with a frequency distribution. Construct the corresponding histogram.

b. Use the histogram in Part (a) to find approximate values of the following percentiles:

i. 50th **iv.** 90th
ii. 70th **v.** 40th
iii. 10th

4.52 The accompanying table gives the mean and standard deviation of reaction times (in seconds) for each of two different stimuli:

	Stimulus	Stimulus
	1	2
Mean	6.0	3.6
Standard deviation	1.2	0.8

If your reaction time is 4.2 seconds for the first stimulus and 1.8 seconds for the second stimulus, to which stimulus are you reacting (compared to other individuals) relatively more quickly?

Bold exercises answered in back ● Data set available online ✦ Video Solution available

4.5 Interpreting and Communicating the Results of Statistical Analyses

As was the case with the graphical displays of Chapter 3, the primary function of the descriptive tools introduced in this chapter is to help us better understand the variables under study. If we have collected data on the amount of money students spend on textbooks at a particular university, most likely we did so because we wanted to learn about the distribution of this variable (amount spent on textbooks) for the population of interest (in this case, students at the university). Numerical measures of center and spread and boxplots help to inform us, and they also allow us to communicate to others what we have learned from the data.

Communicating the Results of Statistical Analyses

When reporting the results of a data analysis, it is common to start with descriptive information about the variables of interest. It is always a good idea to start with a graphical display of the data, and, as we saw in Chapter 3, graphical displays of numerical data are usually described in terms of center, variability, and shape. The numerical measures of this chapter can help you to be more specific in describing the center and spread of a data set.

When describing center and spread, you must first decide which measures to use. Common choices are to use either the sample mean and standard deviation or the sample median and interquartile range (and maybe even a boxplot) to describe center and spread. Because the mean and standard deviation can be sensitive to extreme values in the data set, they are best used when the distribution shape is approximately symmetric and when there are few outliers. If the data set is noticeably skewed or if there are outliers, then the observations are more spread out in one part of the distribution than in the others. In this situation, a five-number summary or a boxplot conveys more information than the mean and standard deviation do.

Interpreting the Results of Statistical Analyses

It is relatively rare to find raw data in published sources. Typically, only a few numerical summary quantities are reported. We must be able to interpret these values and understand what they tell us about the underlying data set.

For example, a university conducted an investigation of the amount of time required to enter the information contained in an application for admission into the university computer system. One of the individuals who performs this task was asked to note starting time and completion time for 50 randomly selected application forms. The resulting entry times (in minutes) were summarized using the mean, median, and standard deviation:

$$\begin{aligned} \bar{x} &= 7.854 \\ \text{median} &= 7.423 \\ s &= 2.129 \end{aligned}$$

What do these summary values tell us about entry times? The average time required to enter admissions data was 7.854 minutes, but the relatively large standard deviation suggests that many observations differ substantially from this mean. The median tells us that half of the applications required less than 7.423 minutes to enter. The fact that the mean exceeds the median suggests that some unusually large values in the data set affected the value of the mean. This last conjecture is confirmed by the stem-and-leaf display of the data given in Figure 4.19.

```
 4 | 8
 5 | 02345679
 6 | 00001234566779
 7 | 223556688
 8 | 23334
 9 | 002
10 | 011168
11 | 134
12 | 2                  Stem:  Ones
13 |                    Leaf:  Tenths
14 | 3
```

FIGURE 4.19
Stem-and-leaf display of data entry times.

The administrators conducting the data-entry study looked at the outlier 14.3 minutes and at the other relatively large values in the data set; they found that the five largest values came from applications that were entered before lunch. After talking with the individual who entered the data, the administrators speculated that morning entry times might differ from afternoon entry times because there tended to be more distractions and interruptions (phone calls, etc.) during the morning hours, when the admissions office generally was busier. When morning and afternoon entry times were separated, the following summary statistics resulted:

Morning (based on $n = 20$ applications):	$\bar{x} = 9.093$	median = 8.743	$s = 2.329$
Afternoon (based on $n = 30$ applications):	$\bar{x} = 7.027$	median = 6.737	$s = 1.529$

Clearly, the average entry time is higher for applications entered in the morning; also, the individual entry times differ more from one another in the mornings than in the afternoons (because the standard deviation for morning entry times, 2.329, is about 1.5 times as large as 1.529, the standard deviation for afternoon entry times).

What to Look for in Published Data

Here are a few questions to ask yourself when you interpret numerical summary measures.

- Is the chosen summary measure appropriate for the type of data collected? In particular, watch for inappropriate use of the mean and standard deviation with categorical data that has simply been coded numerically.
- If both the mean and the median are reported, how do the two values compare? What does this suggest about the distribution of values in the data set? If only the mean or the median was used, was the appropriate measure selected?
- Is the standard deviation large or small? Is the value consistent with your expectations regarding variability? What does the value of the standard deviation tell you about the variable being summarized?
- Can anything of interest be said about the values in the data set by applying Chebyshev's Rule or the Empirical Rule?

For example, consider a study that investigated whether people tend to spend more money when they are paying with a credit card than when they are paying with cash. The authors of the paper "Monopoly Money: The Effect of Payment Coupling and Form on Spending Behavior" (*Journal of Experimental Psychology: Applied* [2008]: 213–225) randomly assigned each of 114 volunteers to one of two experimental groups. Participants were given a menu for a new restaurant that showed nine menu items. They were then asked to estimate the amount they would be willing to pay for each item. A price index was computed for each participant by averaging the nine prices assigned. The difference between the two experimental groups was that the menu viewed by one group showed a credit card logo at the bottom of the menu while there was no credit card logo on the menu that those in the other group viewed. The following passage appeared in the results section of the paper:

> On average, participants were willing to pay more when the credit card logo was present (M = \$4.53, SD = 1.15) than when it was absent (M = \$4.11, SD = 1.06). Thus, even though consumers were not explicitly informed which payment mode they would be using, the mere presence of a credit card logo increased the price that they were willing to pay.

The price index data was also described as mound shaped with no outliers for each of the two groups. Because price index (the average of the prices that a participant assigned to the nine menu items) is a numerical variable, the mean and standard deviation are reasonable measures for summarizing center and spread in the data set. Although the mean for the credit-card-logo group is higher than the mean for the no-logo group, the two standard deviations are similar, indicating similar variability in price index from person to person for the two groups.

Because the distribution of price index values was mound shaped for each of the two groups, we can use the Empirical Rule to tell us a bit more about the distribution. For example, for those in the group who viewed the menu with a credit card logo, approximately 95% of the price index values would have been between

$$4.53 - 2(1.15) = 4.53 - 2.3 = 2.23$$

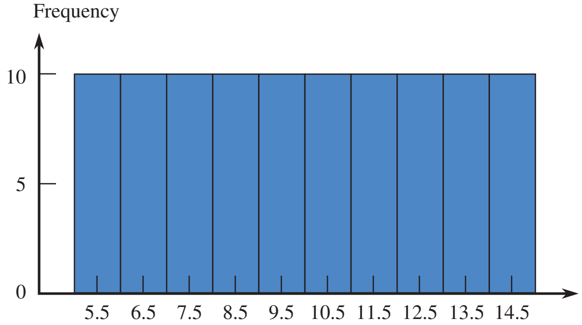
and

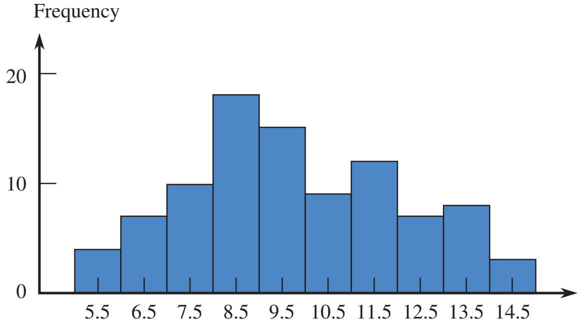
$$4.53 + 2(1.15) = 4.53 + 2.30 = 6.83.$$

A Word to the Wise: Cautions and Limitations

When computing or interpreting numerical descriptive measures, you need to keep in mind the following:

1. Measures of center don't tell all. Although measures of center, such as the mean and the median, do give us a sense of what might be considered a typical value for a variable, this is only one characteristic of a data set. Without additional information about variability and distribution shape, we don't really know much about the behavior of the variable.
2. Data distributions with different shapes can have the same mean and standard deviation. For example, consider the following two histograms:

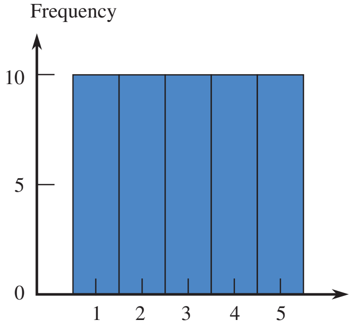




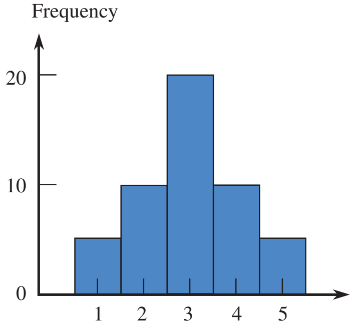
Both histograms summarize data sets that have a mean of 10 and a standard deviation of 2, yet they have different shapes.

3. Both the mean and the standard deviation are sensitive to extreme values in a data set, especially if the sample size is small. If a data distribution is skewed or if the data set has outliers, the median and the interquartile range may be a better choice for describing center and spread.
4. Measures of center and variability describe the values of the variable studied, not the frequencies in a frequency distribution or the heights of the bars in a histogram. For example, consider the following two frequency distributions and histograms:

FREQUENCY DISTRIBUTION A		FREQUENCY DISTRIBUTION B	
Value	Frequency	Value	Frequency
1	10	1	5
2	10	2	10
3	10	3	20
4	10	4	10
5	10	5	5



Histogram A



Histogram B

There is more variability in the data summarized by Frequency Distribution and Histogram A than in the data summarized by Frequency Distribution and Histogram B. This is because the values of the variable described by Histogram and Frequency Distribution B are more concentrated near the mean than are the values for the variable described by Histogram and Frequency Distribution A. Don't be misled by the fact that there is no variability in the frequencies in Frequency Distribution A or the heights of the bars in Histogram A.

5. Be careful with boxplots based on small sample sizes. Boxplots convey information about center, variability, and shape, but when the sample size is small, you should be hesitant to overinterpret shape information. It is really not possible to decide whether a data distribution is symmetric or skewed if only a small sample of observations from the distribution is available.
6. Not all distributions are normal (or even approximately normal). Be cautious in applying the Empirical Rule in situations in which you are not convinced that the data distribution is at least approximately normal. Using the Empirical Rule in such situations can lead to incorrect statements.
7. Watch out for outliers! Unusual observations in a data set often provide important information about the variable under study, so it is important to consider outliers in addition to describing what is typical. Outliers can also be problematic—both because the values of some descriptive measures are influenced by outliers and because some of the methods for drawing conclusions from data may not be appropriate if the data set has outliers.

EXERCISES 4.53 - 4.54

4.53 The authors of the paper "Delayed Time to Defibrillation after In-Hospital Cardiac Arrest" (*New England Journal of Medicine* [2008]: 9–16) described a study of how survival is related to the length of time it takes from the time of a heart attack to the administration of defibrillation therapy. The following is a statement from the paper:

> We identified 6789 patients from 369 hospitals who had in-hospital cardiac arrest due to ventricular fibrillation (69.7%) or pulseless ventricular tachycardia (30.3%). Overall, the median time to defibrillation was 1 minute (interquartile range [was] 3 minutes).

Data from the paper on time to defibrillation (in minutes) for these 6789 patients was used to produce the following Minitab output and boxplot.

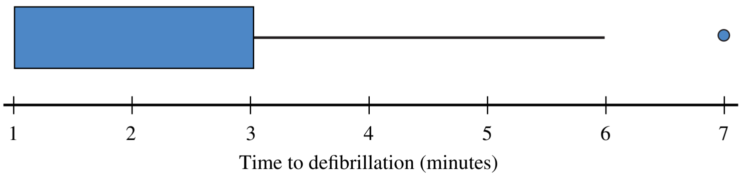
a. Why is there no lower whisker in the given boxplot?

b. How is it possible for the median, the lower quartile, and the minimum value in the data set to all be equal? (Note—this is why you do not see a median line in the box part of the boxplot.)

c. The authors of the paper considered a time to defibrillation of greater than 2 minutes as unacceptable. Based on the given boxplot and summary statistics, is it possible that the percentage of patients having an unacceptable time to defibrillation is greater than 50%? Greater than 25%? Less than 25%? Explain.

```
Descriptive Statistics: Time to Defibrillation

Variable       N      Mean     StDev   Minimum       Q1    Median        Q3   Maximum
Time        6789    2.3737    2.0713    1.0000   1.0000    1.0000    3.0000    7.0000
```



d. Is the outlier shown at 7 a mild outlier or an extreme outlier?

4.54 The paper "Portable Social Groups: Willingness to Communicate, Interpersonal Communication Gratifications, and Cell Phone Use among Young Adults" (*International Journal of Mobile Communications* [2007]: 139–156) describes a study of young adult cell phone use patterns.

a. Comment on the following quote from the paper. Do you agree with the authors?

> Seven sections of an Introduction to Mass Communication course at a large southern university were surveyed in the spring and fall of 2003. The sample was chosen because it offered an excellent representation of the population under study—young adults.

b. Below is another quote from the paper. In this quote, the author reports the mean number of minutes of cell phone use per week for those who participated in the survey. What additional information would have been provided about cell phone use behavior if the author had also reported the standard deviation?

> Based on respondent estimates, users spent an average of 629 minutes (about 10.5 hours) per week using their cell phone on or off line for any reason.

ACTIVITY 4.1 Collecting and Summarizing Numerical Data

In this activity, you will work in groups to collect data that will provide information about how many hours per week, on average, students at your school spend engaged in a particular activity. You will use the sampling plan designed in Activity 2.1 to collect the data.

1. With your group, pick one of the following activities to be the focus of your study:
   i. Surfing the web
   ii. Studying or doing homework
   iii. Watching TV
   iv. Exercising
   v. Sleeping

   or you may choose a different activity, *subject to the approval of your instructor.*
2. Use the plan developed in Activity 2.1 to collect data on the variable you have chosen for your study.
3. Summarize the resulting data using both numerical and graphical summaries. Be sure to address both center and variability.
4. Write a short article for your school paper summarizing your findings regarding student behavior. Your article should include both numerical and graphical summaries.

ACTIVITY 4.2 Airline Passenger Weights

The article "Airlines Should Weigh Passengers, Bags, NTSB Says" (*USA Today*, February 27, 2004) states that the National Transportation Safety Board recommended that airlines weigh passengers and their bags to prevent overloaded planes from attempting to take off. This recommendation was the result of an investigation into the crash of a small commuter plane in 2003, which determined that too much weight contributed to the crash.

Rather than weighing passengers, airlines currently use estimates of average passenger and luggage weights. After the 2003 accident, this estimate was increased by 10 pounds for passengers and 5 pounds for luggage. Although an airplane can fly if it is somewhat overweight if all systems are working properly, if one of the plane's engines fails an overweight plane becomes difficult for the pilot to control.

Assuming that the new estimate of the average passenger weight is accurate, discuss the following questions with a partner and then write a paragraph that answers these questions.

1. What role does variability in passenger weights play in creating a potentially dangerous situation for an airline?
2. Would an airline have a lower risk of a potentially dangerous situation if the variability in passenger weight is large or if it is small?