CHAPTER 3

Data Description

Objectives

After completing this chapter, you should be able to

1. Summarize data using measures of central tendency, such as the mean, median, mode, and midrange.
2. Describe data using measures of variation, such as the range, variance, and standard deviation.
3. Identify the position of a data value in a data set, using various measures of position, such as percentiles, deciles, and quartiles.
4. Use the techniques of exploratory data analysis, including boxplots and five-number summaries, to discover various aspects of data.

Outline

Statistics Today

How Long Are You Delayed by Road Congestion?

No matter where you live, at one time or another, you have been stuck in traffic. To see whether there are more traffic delays in some cities than in others, statisticians make comparisons using descriptive statistics. A statistical study by the Texas Transportation Institute found that a driver is delayed by road congestion an average of 36 hours per year. To see how selected cities compare to this average, see Statistics Today—Revisited at the end of the chapter.

This chapter will show you how to obtain and interpret descriptive statistics such as measures of average, measures of variation, and measures of position.

3–1 Introduction

Chapter 2 showed how one can gain useful information from raw data by organizing them into a frequency distribution and then presenting the data by using various graphs. This chapter shows the statistical methods that can be used to summarize data. The most familiar of these methods is the finding of averages.

For example, one may read that the average speed of a car crossing midtown Manhattan during the day is 5.3 miles per hour or that the average number of minutes an American father of a 4-year-old spends alone with his child each day is 42.[1]

[1] "Harper's Index," *Harper's* magazine.

In the book *American Averages* by Mike Feinsilber and William B. Meed, the authors state:

> *"Average" when you stop to think of it is a funny concept. Although it describes all of us it describes none of us. . . . While none of us wants to be the average American, we all want to know about him or her.*

Interesting Fact

A person has on average 1460 dreams in 1 year.

The authors go on to give examples of averages:

> *The average American man is five feet, nine inches tall; the average woman is five feet, 3.6 inches.*
> *The average American is sick in bed seven days a year missing five days of work.*
> *On the average day, 24 million people receive animal bites.*
> *By his or her 70th birthday, the average American will have eaten 14 steers, 1050 chickens, 3.5 lambs, and 25.2 hogs.*[2]

In these examples, the word *average* is ambiguous, since several different methods can be used to obtain an average. Loosely stated, the average means the center of the distribution or the most typical case. Measures of average are also called *measures of central tendency* and include the *mean, median, mode,* and *midrange.*

Knowing the average of a data set is not enough to describe the data set entirely. Even though a shoe store owner knows that the average size of a man's shoe is size 10, she would not be in business very long if she ordered only size 10 shoes.

As this example shows, in addition to knowing the average, one must know how the data values are dispersed. That is, do the data values cluster around the mean, or are they spread more evenly throughout the distribution? The measures that determine the spread of the data values are called *measures of variation* or *measures of dispersion.* These measures include the *range, variance,* and *standard deviation.*

Finally, another set of measures is necessary to describe data. These measures are called *measures of position.* They tell where a specific data value falls within the data set or its relative position in comparison with other data values. The most common position measures are *percentiles, deciles,* and *quartiles.* These measures are used extensively in psychology and education. Sometimes they are referred to as *norms.*

The measures of central tendency, variation, and position explained in this chapter are part of what is called *traditional statistics.*

Section 3–5 shows the techniques of what is called *exploratory data analysis.* These techniques include the *boxplot* and the *five-number summary.* They can be used to explore data to see what they show (as opposed to the traditional techniques, which are used to confirm conjectures about the data).

3–2 Measures of Central Tendency

Objective 1

Summarize data using measures of central tendency, such as the mean, median, mode, and midrange.

Chapter 1 stated that statisticians use samples taken from populations; however, when populations are small, it is not necessary to use samples since the entire population can be used to gain information. For example, suppose an insurance manager wanted to know the average weekly sales of all the company's representatives. If the company employed a large number of salespeople, say, nationwide, he would have to use a sample and make an inference to the entire sales force. But if the company had only a few salespeople, say, only 87 agents, he would be able to use all representatives' sales for a randomly chosen week and thus use the entire population.

[2] Mike Feinsilber and William B. Meed, *American Averages* (New York: Bantam Doubleday Dell).

Historical Note

In 1796, Adolphe Quetelet investigated the characteristics (heights, weights, etc.) of French conscripts to determine the "average man." Florence Nightingale was so influenced by Quetelet's work that she began collecting and analyzing medical records in the military hospitals during the Crimean War. Based on her work, hospitals began keeping accurate records on their patients.

Measures found by using all the data values in the population are called *parameters*. Measures obtained by using the data values from samples are called *statistics;* hence, the average of the sales from a sample of representatives is called a *statistic,* and the average of sales obtained from the entire population is called a *parameter.*

A **statistic** is a characteristic or measure obtained by using the data values from a sample.

A **parameter** is a characteristic or measure obtained by using all the data values from a specific population.

These concepts as well as the symbols used to represent them will be explained in detail in this chapter.

General Rounding Rule In statistics the basic rounding rule is that when computations are done in the calculation, rounding should not be done until the final answer is calculated. When rounding is done in the intermediate steps, it tends to increase the difference between that answer and the exact one. But in the textbook and solutions manual, it is not practical to show long decimals in the intermediate calculations; hence, the values in the examples are carried out to enough places (usually three or four) to obtain the same answer that a calculator would give after rounding on the last step.

The Mean

The *mean,* also known as the *arithmetic average,* is found by adding the values of the data and dividing by the total number of values. For example, the mean of 3, 2, 6, 5, and 4 is found by adding $3 + 2 + 6 + 5 + 4 = 20$ and dividing by 5; hence, the mean of the data is $20 \div 5 = 4$. The values of the data are represented by X's. In this data set, $X_1 = 3$, $X_2 = 2$, $X_3 = 6$, $X_4 = 5$, and $X_5 = 4$. To show a sum of the total X values, the symbol Σ (the capital Greek letter sigma) is used, and ΣX means to find the sum of the X values in the data set. The summation notation is explained in Appendix A.

The **mean** is the sum of the values, divided by the total number of values. The symbol $\overline{X}$ represents the sample mean.

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} = \frac{\Sigma X}{n}$$

where n represents the total number of values in the sample.

For a population, the Greek letter μ (mu) is used for the mean.

$$\mu = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N} = \frac{\Sigma X}{N}$$

where N represents the total number of values in the population.

In statistics, Greek letters are used to denote parameters, and Roman letters are used to denote statistics. Assume that the data are obtained from samples unless otherwise specified.

Example 3–1

The data represent the number of days off per year for a sample of individuals selected from nine different countries. Find the mean.

20, 26, 40, 36, 23, 42, 35, 24, 30

Source: World Tourism Organization.

Solution

$$\overline{X} = \frac{\Sigma X}{n} = \frac{20 + 26 + 40 + 36 + 23 + 42 + 35 + 24 + 30}{9} = \frac{276}{9} = 30.7 \text{ days}$$

Hence, the mean of the number of days off is 30.7 days.

Example 3–2

The data represent the annual chocolate sales (in billions of dollars) for a sample of seven countries in the world. Find the mean.

2.0, 4.9, 6.5, 2.1, 5.1, 3.2, 16.6

Source: Euromonitor.

Solution

$$\overline{X} = \frac{\Sigma X}{n} = \frac{2.0 + 4.9 + 6.5 + 2.1 + 5.1 + 3.2 + 16.6}{7} = \frac{40.4}{7} = \$5.77 \text{ billion}$$

The mean for the sample is $5.77 billion.

The mean, in most cases, is not an actual data value.

Rounding Rule for the Mean The mean should be rounded to one more decimal place than occurs in the raw data. For example, if the raw data are given in whole numbers, the mean should be rounded to the nearest tenth. If the data are given in tenths, the mean should be rounded to the nearest hundredth, and so on.

The procedure for finding the mean for grouped data uses the midpoints of the classes. This procedure is shown next.

Example 3–3

Using the frequency distribution for Example 2–7, find the mean. The data represent the number of miles run during one week for a sample of 20 runners.

Solution

The procedure for finding the mean for grouped data is given here.

Step 1 Make a table as shown.

A	B	C	D
Class	**Frequency (f)**	**Midpoint (X_m)**	$f \cdot X_m$
5.5–10.5	1		
10.5–15.5	2		
15.5–20.5	3		
20.5–25.5	5		
25.5–30.5	4		
30.5–35.5	3		
35.5–40.5	2		
	$n = 20$		

Step 2 Find the midpoints of each class and enter them in column C.

$$X_m = \frac{5.5 + 10.5}{2} = 8 \qquad \frac{10.5 + 15.5}{2} = 13 \qquad \text{etc.}$$

Interesting Fact

The average time it takes a person to find a new job is 5.9 months.

Step 3 For each class, multiply the frequency by the midpoint, as shown, and place the product in column D.

$1 \cdot 8 = 8 \qquad 2 \cdot 13 = 26 \qquad \text{etc.}$

The completed table is shown here.

A Class	B Frequency (f)	C Midpoint (X_m)	D $f \cdot X_m$
5.5–10.5	1	8	8
10.5–15.5	2	13	26
15.5–20.5	3	18	54
20.5–25.5	5	23	115
25.5–30.5	4	28	112
30.5–35.5	3	33	99
35.5–40.5	2	38	76
	$n = 20$		$\Sigma f \cdot X_m = 490$

Step 4 Find the sum of column D.

Step 5 Divide the sum by n to get the mean.

$$\overline{X} = \frac{\Sigma f \cdot X_m}{n} = \frac{490}{20} = 24.5 \text{ miles}$$

Unusual Stat

A person looks, on average, at about 14 homes before he or she buys one.

The procedure for finding the mean for grouped data assumes that the mean of all the raw data values in each class is equal to the midpoint of the class. In reality, this is not true, since the average of the raw data values in each class usually will not be exactly equal to the midpoint. However, using this procedure will give an acceptable approximation of the mean, since some values fall above the midpoint and other values fall below the midpoint for each class, and the midpoint represents an estimate of all values in the class.

The steps for finding the mean for grouped data are summarized in the next Procedure Table.

Procedure Table

Finding the Mean for Grouped Data

Step 1 Make a table as shown.

A Class	B Frequency (f)	C Midpoint (X_m)	D $f \cdot X_m$

Step 2 Find the midpoints of each class and place them in column C.

Step 3 Multiply the frequency by the midpoint for each class, and place the product in column D.

Step 4 Find the sum of column D.

Step 5 Divide the sum obtained in column D by the sum of the frequencies obtained in column B.

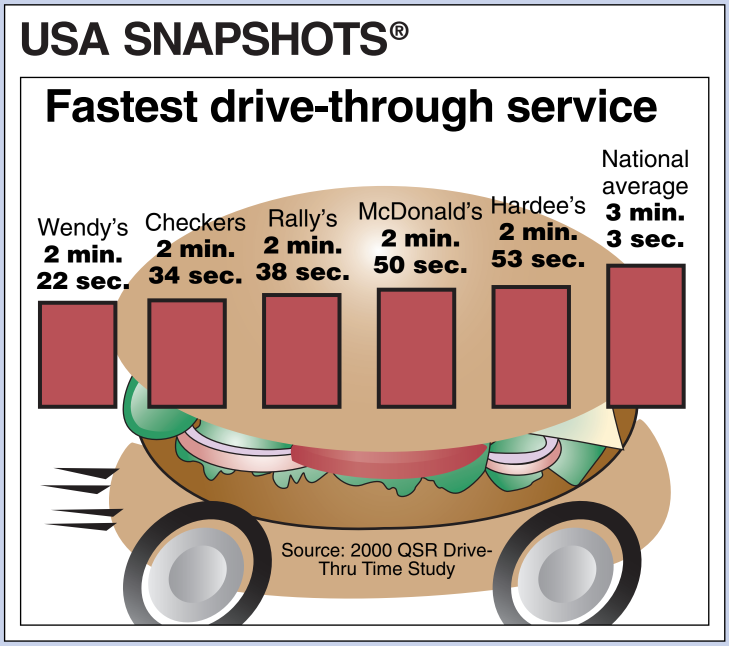
The formula for the mean is

$$\overline{X} = \frac{\Sigma f \cdot X_m}{n}$$

(Note: The symbols $\Sigma f \cdot X_m$ mean to find the sum of the product of the frequency (f) and the midpoint (X_m) for each class.)

Speaking of Statistics

How long do you wait for fast-food service? This Snapshot shows the average times for large fast-food chains. Which type of average (mean, median, or mode) do you think was used?



Source: Copyright 2001, *USA TODAY.* Reprinted with permission.

The Median

An article recently reported that the median income for college professors was $43,250. This measure of central tendency means that one-half of all the professors surveyed earned more than $43,250, and one-half earned less than $43,250.

The *median* is the halfway point in a data set. Before one can find this point, the data must be arranged in order. When the data set is ordered, it is called a **data array.** The median either will be a specific value in the data set or will fall between two values, as shown in Examples 3–4 through 3–8.

The **median** is the midpoint of the data array. The symbol for the median is MD.

Steps in computing the median of a data array

Step 1 Arrange the data in order.

Step 2 Select the middle point.

Example 3–4

The number of rooms in the seven hotels in downtown Pittsburgh is 713, 300, 618, 595, 311, 401, and 292. Find the median.

Source: Interstate Hotels Corporation.

Solution

Step 1 Arrange the data in order.

292, 300, 311, 401, 595, 618, 713

Step 2 Select the middle value.

$$292, 300, 311, \underset{\text{Median}}{\underset{\uparrow}{401}}, 595, 618, 713$$

Hence, the median is 401 rooms.

Example 3–5

Find the median for the ages of seven preschool children. The ages are 1, 3, 4, 2, 3, 5, and 1.

Solution

$$1, 1, 2, \underset{\text{Median}}{\underset{\uparrow}{\boxed{3,}}}\ 3, 4, 5$$

Hence, the median age is 3 years.

Examples 3–4 and 3–5 each had an odd number of values in the data set; hence, the median was an actual data value. When there are an even number of values in the data set, the median will fall between two given values, as illustrated in Examples 3–6, 3–7, and 3–8.

Example 3–6

The number of tornadoes that have occurred in the United States over an 8-year period follows. Find the median.

684, 764, 656, 702, 856, 1133, 1132, 1303

Source: The Universal Almanac.

Solution

$$656, 684, 702, 764\underset{\text{Median}}{\underset{\uparrow}{,}}\ 856, 1132, 1133, 1303$$

Since the middle point falls halfway between 764 and 856, find the median MD by adding the two values and dividing by 2.

$$\text{MD} = \frac{764 + 856}{2} = \frac{1620}{2} = 810$$

The median number of tornadoes is 810.

Example 3–7

The number of cloudy days for the top 10 cloudiest cities is shown. Find the median.

209, 223, 211, 227, 213, 240, 240, 211, 229, 212

Source: National Climatic Data Center.

Solution

Arrange the data in order.

209, 211, 211, 212, 213, 223, 227, 229, 240, 240
↑
Median

$$\text{MD} = \frac{213 + 223}{2} = 218$$

Hence, the median is 218 days.

Example 3–8

Six customers purchased these numbers of magazines: 1, 7, 3, 2, 3, 4. Find the median.

Solution

1, 2, 3, 3, 4, 7
↑
Median

$$\text{MD} = \frac{3 + 3}{2} = 3$$

Hence, the median number of magazines purchased is 3.

The Mode

The third measure of average is called the *mode*. The mode is the value that occurs most often in the data set. It is sometimes said to be the most typical case.

> The value that occurs most often in a data set is called the **mode.**

A data set that has only one value that occurs with the greatest frequency is said to be **unimodal.**

If a data set has two values that occur with the same greatest frequency, both values are considered to be the mode and the data set is said to be **bimodal.** If a data set has more than two values that occur with the same greatest frequency, each value is used as the mode, and the data set is said to be **multimodal.** When no data value occurs more than once, the data set is said to have *no mode.* A data set can have more than one mode or no mode at all. These situations will be shown in some of the examples that follow.

Example 3–9

The following data represent the duration (in days) of U.S. Space Shuttle voyages for the years 1992–1994. Find the mode.

8, 9, 9, 14, 8, 8, 10, 7, 6, 9, 7, 8, 10, 14, 11, 8, 14, 11

Source: The Universal Almanac.

Solution

It is helpful to arrange the data in order, although it is not necessary.

6, 7, 7, 8, 8, 8, 8, 8, 9, 9, 9, 10, 10, 11, 11, 14, 14, 14

Since 8-day voyages occurred 5 times—a frequency larger than any other number—the mode for the data set is 8.

Example 3–10

Find the mode for the number of coal employees per county for 10 selected counties in southwestern Pennsylvania.

110, 731, 1031, 84, 20, 118, 1162, 1977, 103, 752

Source: Pittsburgh Tribune-Review.

Solution

Since each value occurs only once, there is no mode.

Note: Do not say that the mode is zero. That would be incorrect, because in some data, such as temperature, zero can be an actual value.

Example 3–11

The data show the number of licensed nuclear reactors in the United States for a recent 15-year period. Find the mode.

Source: The World Almanac and Book of Facts.

104	104	104	104	104
107	109	109	109	110
109	111	112	111	109

Solution

Since the values 104 and 109 both occur 5 times, the modes are 104 and 109. The data set is said to be bimodal.

The mode for grouped data is the modal class. The **modal class** is the class with the largest frequency.

Example 3–12

Find the modal class for the frequency distribution of miles that 20 runners ran in one week, used in Example 2–7.

Class	Frequency	
5.5–10.5	1	
10.5–15.5	2	
15.5–20.5	3	
20.5–25.5	5	← Modal class
25.5–30.5	4	
30.5–35.5	3	
35.5–40.5	2	

Solution

The modal class is 20.5–25.5, since it has the largest frequency. Sometimes the midpoint of the class is used rather than the boundaries; hence, the mode could also be given as 23 miles per week.

The mode is the only measure of central tendency that can be used in finding the most typical case when the data are nominal or categorical.

Example 3–13

A survey showed this distribution for the number of students enrolled in each field. Find the mode.

Business	1425
Liberal arts	878
Computer science	632
Education	471
General studies	95

Solution

Since the category with the highest frequency is business, the most typical case is a business major.

An extremely high or extremely low data value in a data set can have a striking effect on the mean of the data set. These extreme values are called *outliers*. This is one reason why when analyzing a frequency distribution, you should be aware of any of these values. For the data set shown in Example 3–14, the mean, median, and mode can be quite different because of extreme values. A method for identifying outliers is given in Section 3–4.

Example 3–14

A small company consists of the owner, the manager, the salesperson, and two technicians, all of whose annual salaries are listed here. (Assume that this is the entire population.)

Staff	Salary
Owner	$50,000
Manager	20,000
Salesperson	12,000
Technician	9,000
Technician	9,000

Find the mean, median, and mode.

Solution

$$\mu = \frac{\Sigma X}{N} = \frac{50{,}000 + 20{,}000 + 12{,}000 + 9000 + 9000}{5} = \$20{,}000$$

Hence, the mean is $20,000, the median is $12,000, and the mode is $9000.

In Example 3–14, the mean is much higher than the median or the mode. This is so because the extremely high salary of the owner tends to raise the value of the mean. In this and similar situations, the median should be used as the measure of central tendency.

The Midrange

The *midrange* is a rough estimate of the middle. It is found by adding the lowest and highest values in the data set and dividing by 2. It is a very rough estimate of the average and can be affected by one extremely high or low value.

The **midrange** is defined as the sum of the lowest and highest values in the data set, divided by 2. The symbol MR is used for the midrange.

$$\text{MR} = \frac{\text{lowest value} + \text{highest value}}{2}$$

Example 3–15

In the last two winter seasons, the city of Brownsville, Minnesota, reported these numbers of water-line breaks per month. Find the midrange.

2, 3, 6, 8, 4, 1

Solution

$$\text{MR} = \frac{1 + 8}{2} = \frac{9}{2} = 4.5$$

Hence, the midrange is 4.5.

If the data set contains one extremely large value or one extremely small value, a higher or lower midrange value will result and may not be a typical description of the middle.

Example 3–16

Suppose the number of water-line breaks was as follows: 2, 3, 6, 16, 4, and 1. Find the midrange.

Solution

$$\text{MR} = \frac{1 + 16}{2} = \frac{17}{2} = 8.5$$

Hence, the midrange is 8.5. The value 8.5 is not typical of the average monthly number of breaks, since an excessively high number of breaks, 16, occurred in one month.

In statistics, several measures can be used for an average. The most common measures are the mean, median, mode, and midrange. Each has its own specific purpose and use. Exercises 39 through 41 show examples of other averages, such as the harmonic mean, the geometric mean, and the quadratic mean. Their applications are limited to specific areas, as shown in the exercises.

The Weighted Mean

Sometimes, one must find the mean of a data set in which not all values are equally represented. Consider the case of finding the average cost of a gallon of gasoline for three taxis. Suppose the drivers buy gasoline at three different service stations at a cost of \$2.22, \$2.53, and \$2.63 per gallon. One might try to find the average by using the formula

$$\begin{aligned}\overline{X} &= \frac{\Sigma X}{n} \\ &= \frac{2.22 + 2.53 + 2.63}{3} = \frac{7.38}{3} = \$2.46\end{aligned}$$

But not all drivers purchased the same number of gallons. Hence, to find the true average cost per gallon, one must take into consideration the number of gallons each driver purchased.

The type of mean that considers an additional factor is called the *weighted mean,* and it is used when the values are not all equally represented.

Interesting Fact

The average American drives about 10,000 miles a year.

Find the **weighted mean** of a variable X by multiplying each value by its corresponding weight and dividing the sum of the products by the sum of the weights.

$$\overline{X} = \frac{w_1X_1 + w_2X_2 + \cdots + w_nX_n}{w_1 + w_2 + \cdots + w_n} = \frac{\Sigma wX}{\Sigma w}$$

where $w_1, w_2, \ldots, w_n$ are the weights and $X_1, X_2, \ldots, X_n$ are the values.

Example 3–17 shows how the weighted mean is used to compute a grade point average. Since courses vary in their credit value, the number of credits must be used as weights.

Example 3–17

A student received an A in English Composition I (3 credits), a C in Introduction to Psychology (3 credits), a B in Biology I (4 credits), and a D in Physical Education (2 credits). Assuming A = 4 grade points, B = 3 grade points, C = 2 grade points, D = 1 grade point, and F = 0 grade points, find the student's grade point average.

Solution

Course	Credits (w)	Grade (X)
English Composition I	3	A (4 points)
Introduction to Psychology	3	C (2 points)
Biology I	4	B (3 points)
Physical Education	2	D (1 point)

$$\overline{X} = \frac{\Sigma wX}{\Sigma w} = \frac{3 \cdot 4 + 3 \cdot 2 + 4 \cdot 3 + 2 \cdot 1}{3 + 3 + 4 + 2} = \frac{32}{12} = 2.7$$

The grade point average is 2.7.

Table 3–1 summarizes the measures of central tendency.

Unusual Stat

Of people in the United States, 45% live within 15 minutes of their best friend.

Table 3–1 Summary of Measures of Central Tendency

Measure	Definition	Symbol(s)
Mean	Sum of values, divided by total number of values	$\mu, \overline{X}$
Median	Middle point in data set that has been ordered	MD
Mode	Most frequent data value	None
Midrange	Lowest value plus highest value, divided by 2	MR

Researchers and statisticians must know which measure of central tendency is being used and when to use each measure of central tendency. The properties and uses of the four measures of central tendency are summarized next.

Properties and Uses of Central Tendency

The Mean

1. One computes the mean by using all the values of the data.
2. The mean varies less than the median or mode when samples are taken from the same population and all three measures are computed for these samples.
3. The mean is used in computing other statistics, such as the variance.
4. The mean for the data set is unique and not necessarily one of the data values.
5. The mean cannot be computed for the data in a frequency distribution that has an open-ended class.
6. The mean is affected by extremely high or low values, called outliers, and may not be the appropriate average to use in these situations.

The Median

1. The median is used when one must find the center or middle value of a data set.
2. The median is used when one must determine whether the data values fall into the upper half or lower half of the distribution.
3. The median is used for an open-ended distribution.
4. The median is affected less than the mean by extremely high or extremely low values.

The Mode

1. The mode is used when the most typical case is desired.
2. The mode is the easiest average to compute.
3. The mode can be used when the data are nominal, such as religious preference, gender, or political affiliation.
4. The mode is not always unique. A data set can have more than one mode, or the mode may not exist for a data set.

The Midrange

1. The midrange is easy to compute.
2. The midrange gives the midpoint.
3. The midrange is affected by extremely high or low values in a data set.

Distribution Shapes

Frequency distributions can assume many shapes. The three most important shapes are positively skewed, symmetric, and negatively skewed. Figure 3–1 shows histograms of each.

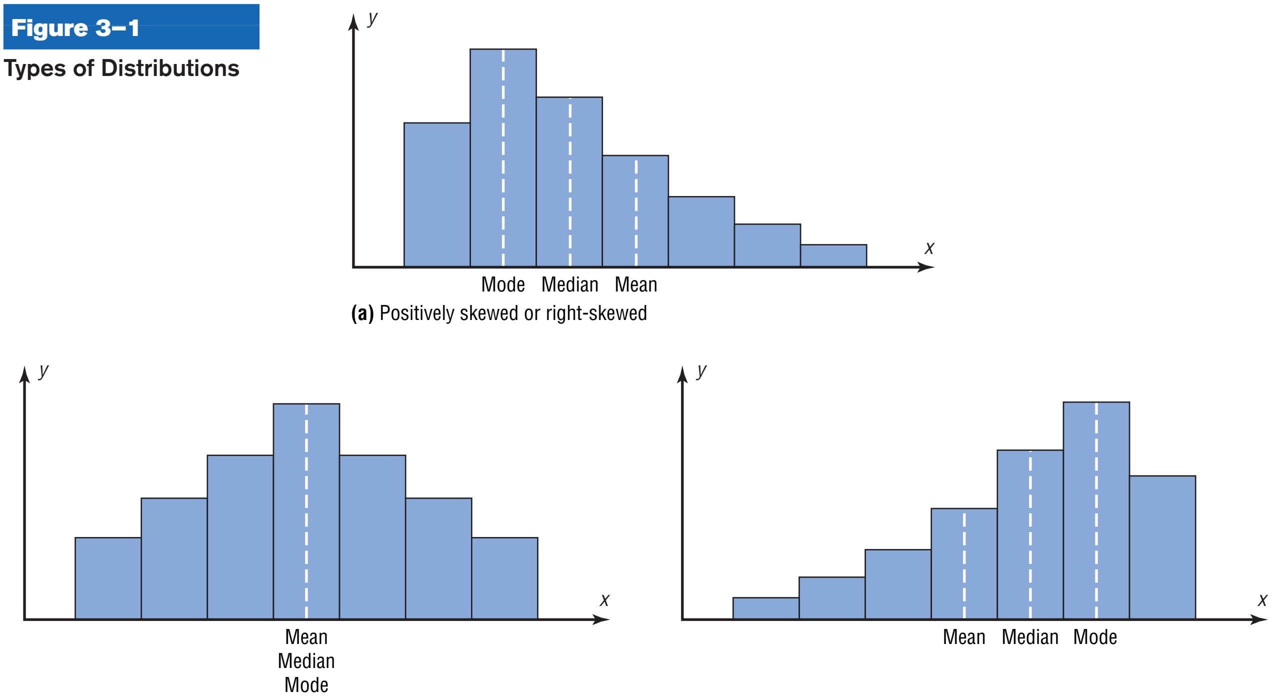
In a **positively skewed** or **right-skewed distribution,** the majority of the data values fall to the left of the mean and cluster at the lower end of the distribution; the “tail” is to the right. Also, the mean is to the right of the median, and the mode is to the left of the median.

For example, if an instructor gave an examination and most of the students did poorly, their scores would tend to cluster on the left side of the distribution. A few high scores would constitute the tail of the distribution, which would be on the right side. Another example of a positively skewed distribution is the incomes of the population of the United States. Most of the incomes cluster about the low end of the distribution; those with high incomes are in the minority and are in the tail at the right of the distribution.

In a **symmetric distribution,** the data values are evenly distributed on both sides of the mean. In addition, when the distribution is unimodal, the mean, median, and mode are the same and are at the center of the distribution. Examples of symmetric distributions are IQ scores and heights of adult males.

Figure 3–1

Types of Distributions



(a) Positively skewed or right-skewed

(b) Symmetric

(c) Negatively skewed or left-skewed

When the majority of the data values fall to the right of the mean and cluster at the upper end of the distribution, with the tail to the left, the distribution is said to be **negatively skewed** or **left-skewed.** Also, the mean is to the left of the median, and the mode is to the right of the median. As an example, a negatively skewed distribution results if the majority of students score very high on an instructor's examination. These scores will tend to cluster to the right of the distribution.

When a distribution is extremely skewed, the value of the mean will be pulled toward the tail, but the majority of the data values will be greater than the mean or less than the mean (depending on which way the data are skewed); hence, the median rather than the mean is a more appropriate measure of central tendency. An extremely skewed distribution can also affect other statistics.

A measure of skewness for a distribution is discussed in Exercise 48 in Section 3–3.

Applying the Concepts 3–2

Teacher Salaries

The following data represent salaries (in dollars) from a school district in Greenwood, South Carolina.

10,000	11,000	11,000	12,500	14,300	17,500
18,000	16,600	19,200	21,560	16,400	107,000

1. First, assume you work for the school board in Greenwood and do not wish to raise taxes to increase salaries. Compute the mean, median, and mode and decide which one would best support your position to not raise salaries.
2. Second, assume you work for the teachers' union and want a raise for the teachers. Use the best measure of central tendency to support your position.
3. Explain how outliers can be used to support one or the other position.

4. If the salaries represented every teacher in the school district, would the averages be parameters or statistics?
5. Which measure of central tendency can be misleading when a data set contains outliers?
6. When one is comparing the measures of central tendency, does the distribution display any skewness? Explain.

See page 170 for the answers.

Exercises 3–2

For Exercises 1 through 8, find (*a*) the mean, (*b*) the median, (*c*) the mode, and (*d*) the midrange.

1. The average undergraduate grade point average (GPA) for the 25 top-ranked medical schools is listed below.

3.80	3.77	3.70	3.74	3.70
3.86	3.76	3.68	3.67	3.57
3.83	3.70	3.80	3.74	3.67
3.78	3.74	3.73	3.65	3.66
3.75	3.64	3.78	3.73	3.64

Source: *U.S. News & World Report Best Graduate Schools.*

2. The heights (in feet) of the 20 highest waterfalls in the world are shown here. (*Note:* The height of Niagara Falls is 182 feet!)

3212	2800	2625	2540	2499	2425	2307	2151	2123	2000
1904	1841	1650	1612	1536	1388	1215	1198	1182	1170

Source: *N.Y. Times Almanac.*

3. The following data are the number of burglaries reported for a specific year for nine western Pennsylvania universities. Which measure of average might be the best in this case? Explain your answer.

61, 11, 1, 3, 2, 30, 18, 3, 7

Source: *Pittsburgh Post Gazette.*

4. For a recent year, the number of suspensions in a sample of public schools in Beaver County was 67, 12, 11, 92, and 13. The number of suspensions in a sample of public schools in Butler County was 56, 12, 18, and 21. Do you think that there is a difference in the averages?

Source: U.S. Department of Education.

5. A researcher claims that each year, there are an average of 300 victims of identity theft in major cities. Twelve were randomly selected, and the number of victims of identity theft in each city is shown. Can you conclude that the researcher was correct?

574	229	663	372	102	88
117	239	465	136	189	75

6. For a recent year, the worth (in billions of dollars) of a sample of the 10 wealthiest people under the age of 60 is shown:

14, 12, 48, 20, 18, 18, 12.6, 10.4, 7.3, 5.3

The worth of a sample of 10 of the wealthiest people age 60 and over is shown:

41.0, 13.7, 10.0, 18.0, 11.3, 7.6, 7, 18, 20, 60

Based on the averages, can you conclude that those 60 years old and over have a higher net worth?

Source: *Forbes* magazine.

7. Twelve major earthquakes had Richter magnitudes shown here.

7.0, 6.2, 7.7, 8.0, 6.4, 6.2,
7.2, 5.4, 6.4, 6.5, 7.2, 5.4

Which would you consider the best measure of average?

Source: *The Universal Almanac*.

8. The data shown are the total compensation (in millions of dollars) for the 50 top-paid CEOs for a recent year. Compare the averages, and state which one you think is the best measure.

17.5	18.0	36.8	31.7	31.7
17.3	24.3	47.7	38.5	17.0
23.7	16.5	25.1	17.4	18.0
37.6	19.7	21.4	28.6	21.6
19.3	20.0	16.9	25.2	19.8
25.0	17.2	20.4	20.1	29.1
19.1	25.2	23.2	25.9	24.0
41.7	24.0	16.8	26.8	31.4
16.9	17.2	24.1	35.2	19.1
22.9	18.2	25.4	35.4	25.5

Source: *USA TODAY*.

9. Find the (*a*) mean, (*b*) median, (*c*) mode, and (*d*) midrange for the data in Exercise 17 in Section 2–2. Is the distribution symmetric or skewed? Use the individual data values.

10. Find the (*a*) mean, (*b*) median, (*c*) mode, and (*d*) midrange for the distances of the home runs for McGwire and Sosa, using the data in Exercise 18 in Section 2–2.

Compare the means. Decide if the means are approximately equal or if one of the players is hitting longer home runs. Use the individual data values.

11. These data represent the number of traffic fatalities for two specific years for 27 selected states. Find the (*a*) mean, (*b*) median, (*c*) mode, and (*d*) midrange for each data set. Are the four measures of average for fatalities for year 1 the same as those for year 2? (The data in this exercise will be used in Exercise 15 in Section 3–3.)

Year 1			Year 2		
1113	1488	868	1100	260	205
1031	262	1109	970	1430	300
4192	1586	215	4040	460	350
645	527	254	620	480	485
121	442	313	125	405	85
2805	444	485	2805	690	1430
900	653	170	1555	1160	70
74	1480	69	180	3360	325
158	3181	326	875	705	145

Source: *USA TODAY*.

For Exercises 12 through 21, find the (*a*) mean and (*b*) modal class.

12. For 108 randomly selected college students, this exam score frequency distribution was obtained. (The data in this exercise will be used in Exercise 18 in Section 3–3.)

Class limits	Frequency
90–98	6
99–107	22
108–116	43
117–125	28
126–134	9

13. The scores for the LPGA—Giant Eagle were

Score	Frequency
202–204	2
205–207	7
208–210	16
211–213	26
214–216	18
217–219	4

Source: www.LPGA.com

14. Thirty automobiles were tested for fuel efficiency (in miles per gallon). This frequency distribution was obtained. (The data in this exercise will be used in Exercise 20 in Section 3–3.)

Class boundaries	Frequency
7.5–12.5	3
12.5–17.5	5
17.5–22.5	15
22.5–27.5	5
27.5–32.5	2

15. The data show the number of murders in 25 selected cities in a given state.

Number	Frequency
34–96	13
97–159	2
160–222	0
223–285	5
286–348	1
349–411	1
412–474	0
475–537	1
538–600	2

Do you think that the mean is the best measure of average for these data? Explain your answer. (The information in the exercise will be used in Exercise 21 in Section 3–3.)

16. Find the mean and modal class for the two frequency distributions in Exercises 8 and 18 in Section 2–3. Are the "average" reactions the same? Explain your answer.

17. Eighty randomly selected lightbulbs were tested to determine their lifetimes (in hours). This frequency distribution was obtained. (The data in this exercise will be used in Exercise 23 in Section 3–3.)

Class boundaries	Frequency
52.5–63.5	6
63.5–74.5	12
74.5–85.5	25
85.5–96.5	18
96.5–107.5	14
107.5–118.5	5

18. These data represent the net worth (in millions of dollars) of 45 national corporations.

Class limits	Frequency
10–20	2
21–31	8
32–42	15
43–53	7
54–64	10
65–75	3

19. The cost per load (in cents) of 35 laundry detergents tested by a consumer organization is shown. (The data in this exercise will be used for Exercise 19 in Section 3–3.)

Class limits	Frequency
13–19	2
20–26	7
27–33	12
34–40	5
41–47	6
48–54	1
55–61	0
62–68	2

20. This frequency distribution represents the commission earned (in dollars) by 100 salespeople employed at several branches of a large chain store.

Class limits	Frequency
150–158	5
159–167	16
168–176	20
177–185	21
186–194	20
195–203	15
204–212	3

21. This frequency distribution represents the data obtained from a sample of 75 copying machine service technicians. The values represent the days between service calls for various copying machines.

Class boundaries	Frequency
15.5–18.5	14
18.5–21.5	12
21.5–24.5	18
24.5–27.5	10
27.5–30.5	15
30.5–33.5	6

22. Find the mean and modal class for the data in Exercise 12 in Section 2–2.

23. Find the mean and modal class for the data in Exercise 13 in Section 2–2.

24. Find the mean and modal class for the data in Exercise 14 in Section 2–2.

25. Find the mean and modal class for the data in Exercise 15 in Section 2–2.

26. Find the weighted mean price of three models of automobiles sold. The number and price of each model sold are shown in this list.

Model	Number	Price
A	8	$10,000
B	10	12,000
C	12	8,000

27. Using the weighted mean, find the average number of grams of fat per ounce of meat or fish that a person would consume over a 5-day period if he ate these:

Meat or fish	Fat (g/oz)
3 oz fried shrimp	3.33
3 oz veal cutlet (broiled)	3.00
2 oz roast beef (lean)	2.50
2.5 oz fried chicken drumstick	4.40
4 oz tuna (canned in oil)	1.75

Source: *The World Almanac and Book of Facts.*

28. A recent survey of a new diet cola reported the following percentages of people who liked the taste. Find the weighted mean of the percentages.

Area	% Favored	Number surveyed
1	40	1000
2	30	3000
3	50	800

29. The costs of three models of helicopters are shown here. Find the weighted mean of the costs of the models.

Model	Number sold	Cost
Sunscraper	9	$427,000
Skycoaster	6	365,000
High-flyer	12	725,000

30. An instructor grades exams, 20%; term paper, 30%; final exam, 50%. A student had grades of 83, 72, and 90, respectively, for exams, term paper, and final exam. Find the student's final average. Use the weighted mean.

31. Another instructor gives four 1-hour exams and one final exam, which counts as two 1-hour exams. Find a student's grade if she received 62, 83, 97, and 90 on the 1-hour exams and 82 on the final exam.

32. For these situations, state which measure of central tendency—mean, median, or mode—should be used.

a. The most typical case is desired.

b. The distribution is open-ended.

c. There is an extreme value in the data set.

d. The data are categorical.

e. Further statistical computations will be needed.

f. The values are to be divided into two approximately equal groups, one group containing the larger values and one containing the smaller values.

33. Describe which measure of central tendency—mean, median, or mode—was probably used in each situation.

a. One-half of the factory workers make more than $5.37 per hour, and one-half make less than $5.37 per hour.

b. The average number of children per family in the Plaza Heights Complex is 1.8.

c. Most people prefer red convertibles over any other color.

d. The average person cuts the lawn once a week.

e. The most common fear today is fear of speaking in public.

f. The average age of college professors is 42.3 years.

34. What types of symbols are used to represent sample statistics? Give an example. What types of symbols are used to represent population parameters? Give an example.

35. A local fast-food company claims that the average salary of its employees is $13.23 per hour. An employee states that most employees make minimum wage. If both are being truthful, how could both be correct?

Extending the Concepts

36. If the mean of five values is 64, find the sum of the values.

37. If the mean of five values is 8.2 and four of the values are 6, 10, 7, and 12, find the fifth value.

38. Find the mean of 10, 20, 30, 40, and 50.

a. Add 10 to each value and find the mean.

b. Subtract 10 from each value and find the mean.

c. Multiply each value by 10 and find the mean.

d. Divide each value by 10 and find the mean.

e. Make a general statement about each situation.

39. The *harmonic mean* (HM) is defined as the number of values, divided by the sum of the reciprocals of each value. The formula is

$$\text{HM} = \frac{n}{\Sigma(1/X)}$$

For example, the harmonic mean of 1, 4, 5, and 2 is

$$\text{HM} = \frac{4}{1/1 + 1/4 + 1/5 + 1/2} = 2.05$$

This mean is useful for finding the average speed. Suppose a person drove 100 miles at 40 miles per hour and returned, driving 50 miles per hour. The average miles per hour is *not* 45 miles per hour, which is found by adding 40 and 50 and dividing by 2. The average is found as shown.

Since

$$\text{Time} = \text{distance} \div \text{rate}$$

then

$$\text{Time 1} = \frac{100}{40} = 2.5 \text{ hours to make the trip}$$

$$\text{Time 2} = \frac{100}{50} = 2 \text{ hours to return}$$

Hence, the total time is 4.5 hours, and the total miles driven are 200. Now, the average speed is

$$\text{Rate} = \frac{\text{distance}}{\text{time}} = \frac{200}{4.5} = 44.44 \text{ miles per hour}$$

This value can also be found by using the harmonic mean formula

$$\text{HM} = \frac{2}{1/40 + 1/50} = 44.44$$

Using the harmonic mean, find each of these.

a. A salesperson drives 300 miles round trip at 30 miles per hour going to Chicago and 45 miles per hour returning home. Find the average miles per hour.

b. A bus driver drives the 50 miles to West Chester at 40 miles per hour and returns, driving 25 miles per hour. Find the average miles per hour.

c. A carpenter buys $500 worth of nails at $50 per pound and $500 worth of nails at $10 per pound. Find the average cost of 1 pound of nails.

40. The *geometric mean* (GM) is defined as the *n*th root of the product of *n* values. The formula is

$$\text{GM} = \sqrt[n]{(X_1)(X_2)(X_3)\cdots(X_n)}$$

The geometric mean of 4 and 16 is

$$\text{GM} = \sqrt{(4)(16)} = \sqrt{64} = 8$$

The geometric mean of 1, 3, and 9 is

$$\text{GM} = \sqrt[3]{(1)(3)(9)} = \sqrt[3]{27} = 3$$

The geometric mean is useful in finding the average of percentages, ratios, indexes, or growth rates. For example, if a person receives a 20% raise after 1 year of service and a 10% raise after the second year of service, the average percentage raise per year is not 15 but 14.89%, as shown.

$$GM = \sqrt{(1.2)(1.1)} = 1.1489$$

or

$$GM = \sqrt{(120)(110)} = 114.89\%$$

His salary is 120% at the end of the first year and 110% at the end of the second year. This is equivalent to an average of 14.89%, since 114.89% − 100% = 14.89%.

This answer can also be shown by assuming that the person makes $10,000 to start and receives two raises of 20% and 10%.

$$\text{Raise } 1 = 10{,}000 \cdot 20\% = \$2000$$
$$\text{Raise } 2 = 12{,}000 \cdot 10\% = \$1200$$

His total salary raise is $3200. This total is equivalent to

$$\begin{aligned} \$10{,}000 \cdot 14.89\% &= \$1489.00 \\ \$11{,}489 \cdot 14.89\% &= \underline{1710.71} \\ & \$3199.71 \approx \$3200 \end{aligned}$$

Find the geometric mean of each of these.

a. The growth rates of the Living Life Insurance Corporation for the past 3 years were 35, 24, and 18%.

b. A person received these percentage raises in salary over a 4-year period: 8, 6, 4, and 5%.

c. A stock increased each year for 5 years at these percentages: 10, 8, 12, 9, and 3%.

d. The price increases, in percentages, for the cost of food in a specific geographic region for the past 3 years were 1, 3, and 5.5%.

41. A useful mean in the physical sciences (such as voltage) is the *quadratic mean* (QM), which is found by taking the square root of the average of the squares of each value. The formula is

$$QM = \sqrt{\frac{\Sigma X^2}{n}}$$

The quadratic mean of 3, 5, 6, and 10 is

$$QM = \sqrt{\frac{3^2 + 5^2 + 6^2 + 10^2}{4}}$$
$$= \sqrt{42.5} = 6.52$$

Find the quadratic mean of 8, 6, 3, 5, and 4.

42. An approximate median can be found for data that have been grouped into a frequency distribution. First it is necessary to find the median class. This is the class that contains the median value. That is the $n/2$ data value. Then it is assumed that the data values are evenly distributed throughout the median class. The formula is

$$MD = \frac{(n/2) - \text{cf}}{f}(w) + L_m$$

where

n = sum of frequencies
cf = cumulative frequency of class immediately preceding the median class
w = width of median class
f = frequency of median class
L_m = lower boundary of median class

Using this formula, find the median for data in the frequency distribution of Exercise 15.

Technology *Step by Step*

Excel
Step by Step

Finding the Central Tendency

Example XL3–1

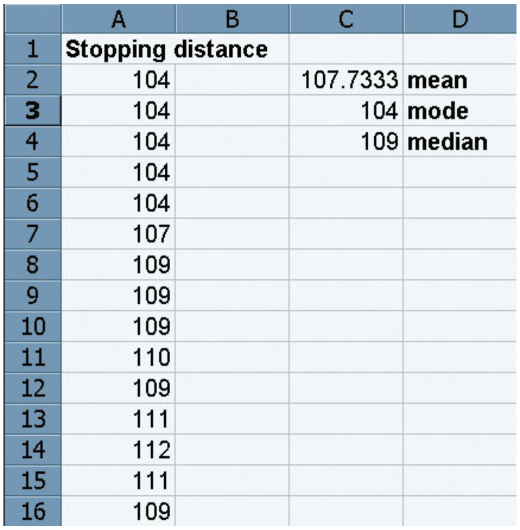
To find the mean, mode, and median of a data set:

1. Enter the numbers in a range of cells (here shown as the numbers in cells A2 to A16). We use the data from Example 3–11 on licensed nuclear reactors:

   104, 104, 104, 104, 104, 107, 109, 109, 109, 110, 109, 111, 112, 111, 109

2. For the mean, enter **=AVERAGE (A2:A16)** in a blank cell.
3. For the mode, enter **=MODE (A2:A16)** in a blank cell.
4. For the median, enter **=MEDIAN (A2:A16)** in a blank cell.

These three functions are available from the standard toolbar by clicking the f_x icon and scrolling down the list of statistical functions. *Note:* For distributions that are *bimodal,* like this one, the Excel MODE function reports the first mode only. A better practice is to use the Histogram routine from the Data Analysis Add-in, which reports actual counts in a table.



	A	B	C	D
1	Stopping distance			
2	104		107.7333	mean
3	104		104	mode
4	104		109	median
5	104			
6	104			
7	107			
8	109			
9	109			
10	109			
11	110			
12	109			
13	111			
14	112			
15	111			
16	109			

3–3 Measures of Variation

In statistics, to describe the data set accurately, statisticians must know more than the measures of central tendency. Consider Example 3–18.

Example 3–18

Objective 2

Describe data using measures of variation, such as the range, variance, and standard deviation.

A testing lab wishes to test two experimental brands of outdoor paint to see how long each will last before fading. The testing lab makes 6 gallons of each paint to test. Since different chemical agents are added to each group and only six cans are involved, these two groups constitute two small populations. The results (in months) are shown. Find the mean of each group.

Brand A	Brand B
10	35
60	45
50	30
30	35
40	40
20	25

Solution

The mean for brand A is

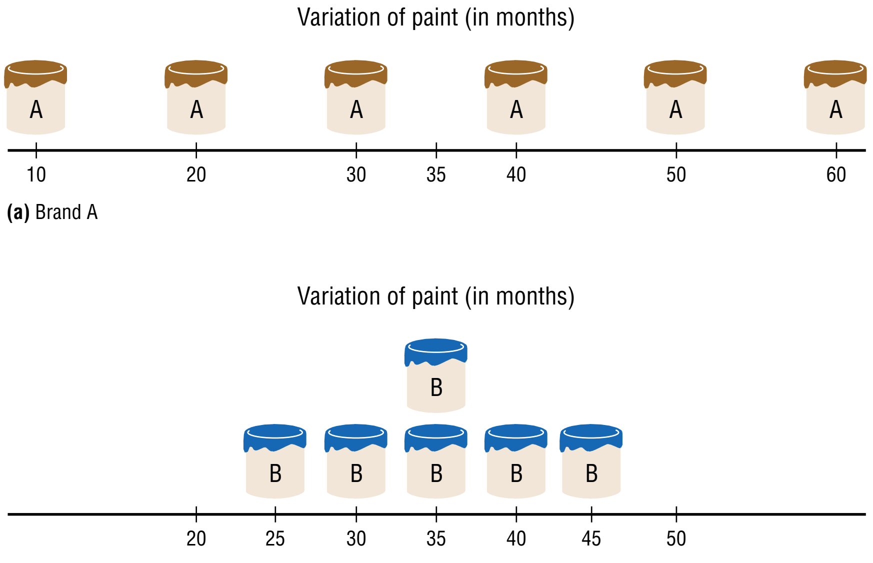
$$\mu = \frac{\Sigma X}{N} = \frac{210}{6} = 35 \text{ months}$$

The mean for brand B is

$$\mu = \frac{\Sigma X}{N} = \frac{210}{6} = 35 \text{ months}$$

Figure 3–2

Examining Data Sets Graphically



(a) Brand A

(b) Brand B

Since the means are equal in Example 3–18, one might conclude that both brands of paint last equally well. However, when the data sets are examined graphically, a somewhat different conclusion might be drawn. See Figure 3–2.

As Figure 3–2 shows, even though the means are the same for both brands, the spread, or variation, is quite different. Figure 3–2 shows that brand B performs more consistently; it is less variable. For the spread or variability of a data set, three measures are commonly used: *range, variance,* and *standard deviation.* Each measure will be discussed in this section.

Range

The range is the simplest of the three measures and is defined now.

The **range** is the highest value minus the lowest value. The symbol R is used for the range.

$$R = \text{highest value} - \text{lowest value}$$

Example 3–19

Find the ranges for the paints in Example 3–18.

Solution

For brand A, the range is

$$R = 60 - 10 = 50 \text{ months}$$

For brand B, the range is

$$R = 45 - 25 = 20 \text{ months}$$

Make sure the range is given as a single number.

The range for brand A shows that 50 months separate the largest data value from the smallest data value. For brand B, 20 months separate the largest data value from the smallest data value, which is less than one-half of brand A's range.

One extremely high or one extremely low data value can affect the range markedly, as shown in Example 3–20.

Example 3–20

The salaries for the staff of the XYZ Manufacturing Co. are shown here. Find the range.

Staff	Salary
Owner	$100,000
Manager	40,000
Sales representative	30,000
Workers	25,000
	15,000
	18,000

Solution

The range is $R = \$100{,}000 - \$15{,}000 = \$85{,}000$.

Since the owner's salary is included in the data for Example 3–20, the range is a large number. To have a more meaningful statistic to measure the variability, statisticians use measures called the *variance* and *standard deviation.*

Population Variance and Standard Deviation

Before the variance and standard deviation are defined formally, the computational procedure will be shown, since the definition is derived from the procedure.

Rounding Rule for the Standard Deviation The rounding rule for the standard deviation is the same as that for the mean. The final answer should be rounded to one more decimal place than that of the original data.

Example 3–21

Find the variance and standard deviation for the data set for brand A paint in Example 3–18.

10, 60, 50, 30, 40, 20

Solution

Step 1 Find the mean for the data.

$$\mu = \frac{\Sigma X}{N} = \frac{10 + 60 + 50 + 30 + 40 + 20}{6} = \frac{210}{6} = 35$$

Step 2 Subtract the mean from each data value.

$$10 - 35 = -25 \qquad 50 - 35 = +15 \qquad 40 - 35 = +5$$
$$60 - 35 = +25 \qquad 30 - 35 = -5 \qquad 20 - 35 = -15$$

Step 3 Square each result.

$$(-25)^2 = 625 \qquad (+15)^2 = 225 \qquad (+5)^2 = 25$$
$$(+25)^2 = 625 \qquad (-5)^2 = 25 \qquad (-15)^2 = 225$$

Step 4 Find the sum of the squares.

$$625 + 625 + 225 + 25 + 25 + 225 = 1750$$

Step 5 Divide the sum by N to get the variance.

Variance = 1750 ÷ 6 = 291.7

Step 6 Take the square root of the variance to get the standard deviation. Hence, the standard deviation equals $\sqrt{291.7}$, or 17.1. It is helpful to make a table.

A Values (X)	B $X - \mu$	C $(X - \mu)^2$
10	−25	625
60	+25	625
50	+15	225
30	−5	25
40	+5	25
20	−15	225
		1750

Column A contains the raw data X. Column B contains the differences $X - \mu$ obtained in step 2. Column C contains the squares of the differences obtained in step 3.

The preceding computational procedure reveals several things. First, the square root of the variance gives the standard deviation; and vice versa, squaring the standard deviation gives the variance. Second, the variance is actually the average of the square of the distance that each value is from the mean. Therefore, if the values are near the mean, the variance will be small. In contrast, if the values are far from the mean, the variance will be large.

One might wonder why the squared distances are used instead of the actual distances. One reason is that the sum of the distances will always be zero. To verify this result for a specific case, add the values in column B of the table in Example 3–21. When each value is squared, the negative signs are eliminated.

Finally, why is it necessary to take the square root? The reason is that since the distances were squared, the units of the resultant numbers are the squares of the units of the original raw data. Finding the square root of the variance puts the standard deviation in the same units as the raw data.

When you are finding the square root, always use its positive or principal value, since the variance and standard deviation of a data set can never be negative.

The **variance** is the average of the squares of the distance each value is from the mean. The symbol for the population variance is σ^2 (σ is the Greek lowercase letter sigma).

The formula for the population variance is

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N}$$

where

X = individual value
μ = population mean
N = population size

The **standard deviation** is the square root of the variance. The symbol for the population standard deviation is σ.

The corresponding formula for the population standard deviation is

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(X - \mu)^2}{N}}$$

Example 3–22

Find the variance and standard deviation for brand B paint data in Example 3–18. The months were

35, 45, 30, 35, 40, 25

Solution

Step 1 Find the mean.

$$\mu = \frac{\Sigma X}{N} = \frac{35 + 45 + 30 + 35 + 40 + 25}{6} = \frac{210}{6} = 35$$

Step 2 Subtract the mean from each value, and place the result in column B of the table.

Step 3 Square each result and place the squares in column C of the table.

A	B	C
X	$X - \mu$	$(X - \mu)^2$
35	0	0
45	10	100
30	−5	25
35	0	0
40	5	25
25	−10	100

Step 4 Find the sum of the squares in column C.

$$\Sigma(X - \mu)^2 = 0 + 100 + 25 + 0 + 25 + 100 = 250$$

Step 5 Divide the sum by N to get the variance.

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N} = \frac{250}{6} = 41.7$$

Step 6 Take the square root to get the standard deviation.

$$\sigma = \sqrt{\frac{\Sigma(X - \mu)^2}{N}} = \sqrt{41.7} = 6.5$$

Hence, the standard deviation is 6.5.

Interesting Fact

Each person receives on average 598 pieces of mail per year.

Since the standard deviation of brand A is 17.1 (see Example 3–21) and the standard deviation of brand B is 6.5, the data are more variable for brand A. *In summary, when the means are equal, the larger the variance or standard deviation is, the more variable the data are.*

Sample Variance and Standard Deviation

When computing the variance for a sample, one might expect the following expression to be used:

$$\frac{\Sigma(X - \overline{X})^2}{n}$$

where $\overline{X}$ is the sample mean and n is the sample size. *This formula is not usually used, however, since in most cases the purpose of calculating the statistic is to estimate the* *corresponding parameter.* For example, the sample mean $\overline{X}$ is used to estimate the population mean μ. The expression

$$\frac{\Sigma(X-\overline{X})^2}{n}$$

does not give the best estimate of the population variance because when the population is large and the sample is small (usually less than 30), the variance computed by this formula usually underestimates the population variance. Therefore, instead of dividing by n, find the variance of the sample by dividing by $n-1$, giving a slightly larger value and an *unbiased* estimate of the population variance.

The formula for the sample variance, denoted by s^2, is

$$s^2=\frac{\Sigma(X-\overline{X})^2}{n-1}$$

where
$\overline{X}$ = sample mean
n = sample size

To find the standard deviation of a sample, one must take the square root of the sample variance, which was found by using the preceding formula.

Formula for the Sample Standard Deviation

The standard deviation of a sample (denoted by s) is

$$s=\sqrt{s^2}=\sqrt{\frac{\Sigma(X-\overline{X})^2}{n-1}}$$

where
X = individual value
$\overline{X}$ = sample mean
n = sample size

Shortcut formulas for computing the variance and standard deviation are presented next and will be used in the remainder of the chapter and in the exercises. These formulas are mathematically equivalent to the preceding formulas and do not involve using the mean. They save time when repeated subtracting and squaring occur in the original formulas. They are also more accurate when the mean has been rounded.

Shortcut or Computational Formulas for s^2 and s

The shortcut formulas for computing the variance and standard deviation for data obtained from samples are as follows.

Variance	**Standard deviation**
$s^2=\frac{\Sigma X^2-[(\Sigma X)^2/n]}{n-1}$	$s=\sqrt{\frac{\Sigma X^2-[(\Sigma X)^2/n]}{n-1}}$

Examples 3–23 and 3–24 explain how to use the shortcut formulas.

Example 3–23

Find the sample variance and standard deviation for the amount of European auto sales for a sample of 6 years shown. The data are in millions of dollars.

11.2, 11.9, 12.0, 12.8, 13.4, 14.3

Source: USA TODAY.

Solution

Step 1 Find the sum of the values.

$$\Sigma X = 11.2 + 11.9 + 12.0 + 12.8 + 13.4 + 14.3 = 75.6$$

Step 2 Square each value and find the sum.

$$\Sigma X^2 = 11.2^2 + 11.9^2 + 12.0^2 + 12.8^2 + 13.4^2 + 14.3^2 = 958.94$$

Step 3 Substitute in the formulas and solve.

$$s^2 = \frac{\Sigma X^2 - [(\Sigma X)^2/n]}{n - 1} = \frac{958.94 - [(75.6)^2/6]}{5}$$

$$= 1.28$$

The variance of the sample is 1.28.

$$s = \sqrt{1.28} = 1.13$$

Hence, the sample standard deviation is 1.13.

Note that ΣX^2 is not the same as $(\Sigma X)^2$. The notation ΣX^2 means to square the values first, then sum; $(\Sigma X)^2$ means to sum the values first, then square the sum.

Variance and Standard Deviation for Grouped Data

The procedure for finding the variance and standard deviation for grouped data is similar to that for finding the mean for grouped data, and it uses the midpoints of each class.

Example 3–24

Find the variance and the standard deviation for the frequency distribution of the data in Example 2–7. The data represent the number of miles that 20 runners ran during one week.

Class	Frequency	Midpoint
5.5–10.5	1	8
10.5–15.5	2	13
15.5–20.5	3	18
20.5–25.5	5	23
25.5–30.5	4	28
30.5–35.5	3	33
35.5–40.5	2	38

Solution

Step 1 Make a table as shown, and find the midpoint of each class.

A Class	B Frequency (f)	C Midpoint (X_m)	D $f \cdot X_m$	E $f \cdot X_m^2$
5.5–10.5	1	8		
10.5–15.5	2	13		
15.5–20.5	3	18		
20.5–25.5	5	23		
25.5–30.5	4	28		
30.5–35.5	3	33		
35.5–40.5	2	38		

Step 2 Multiply the frequency by the midpoint for each class, and place the products in column D.

$$1 \cdot 8 = 8 \qquad 2 \cdot 13 = 26 \qquad \ldots \qquad 2 \cdot 38 = 76$$

Step 3 Multiply the frequency by the square of the midpoint, and place the products in column E.

$$1 \cdot 8^2 = 64 \qquad 2 \cdot 13^2 = 338 \qquad \ldots \qquad 2 \cdot 38^2 = 2888$$

Step 4 Find the sums of columns B, D, and E. The sum of column B is n, the sum of column D is $\Sigma f \cdot X_m$, and the sum of column E is $\Sigma f \cdot X_m^2$. The completed table is shown.

A Class	B Frequency	C Midpoint	D $f \cdot X_m$	E $f \cdot X_m^2$
5.5–10.5	1	8	8	64
10.5–15.5	2	13	26	338
15.5–20.5	3	18	54	972
20.5–25.5	5	23	115	2,645
25.5–30.5	4	28	112	3,136
30.5–35.5	3	33	99	3,267
35.5–40.5	2	38	76	2,888
	$n = 20$		$\Sigma f \cdot X_m = 490$	$\Sigma f \cdot X_m^2 = 13{,}310$

Step 5 Substitute in the formula and solve for s^2 to get the variance.

$$s^2 = \frac{\Sigma f \cdot X_m^2 - [(\Sigma f \cdot X_m)^2/n]}{n-1}$$

$$= \frac{13{,}310 - [(490)^2/20]}{20-1} = 68.7$$

Step 6 Take the square root to get the standard deviation.

$$s = \sqrt{68.7} = 8.3$$

Unusual Stat

At birth men outnumber women by 2%. By age 25, the number of men living is about equal to the number of women living. By age 65, there are 14% more women living than men.

Be sure to use the number found in the sum of column B (i.e., the sum of the frequencies) for n. Do not use the number of classes.

The steps for finding the variance and standard deviation for grouped data are summarized in this Procedure Table.

Procedure Table

Finding the Sample Variance and Standard Deviation for Grouped Data

Step 1 Make a table as shown, and find the midpoint of each class.

A	B	C	D	E
Class	Frequency	Midpoint	$f \cdot X_m$	$f \cdot X_m^2$

Step 2 Multiply the frequency by the midpoint for each class, and place the products in column D.

Step 3 Multiply the frequency by the square of the midpoint, and place the products in column E.

Step 4 Find the sums of columns B, D, and E. (The sum of column B is n. The sum of column D is $\Sigma f \cdot X_m$. The sum of column E is $\Sigma f \cdot X_m^2$.)

Step 5 Substitute in the formula and solve to get the variance.

$$s^2 = \frac{\Sigma f \cdot X_m^2 - [(\Sigma f \cdot X_m)^2/n]}{n - 1}$$

Step 6 Take the square root to get the standard deviation.

The three measures of variation are summarized in Table 3–2.

Unusual Stat

The average number of times that a man cries in a month is 1.4.

Table 3–2 Summary of Measures of Variation

Measure	Definition	Symbol(s)
Range	Distance between highest value and lowest value	R
Variance	Average of the squares of the distance that each value is from the mean	σ^2, s^2
Standard deviation	Square root of the variance	σ, s

Uses of the Variance and Standard Deviation

1. As previously stated, variances and standard deviations can be used to determine the spread of the data. If the variance or standard deviation is large, the data are more dispersed. This information is useful in comparing two (or more) data sets to determine which is more (most) variable.
2. The measures of variance and standard deviation are used to determine the consistency of a variable. For example, in the manufacture of fittings, such as nuts and bolts, the variation in the diameters must be small, or the parts will not fit together.
3. The variance and standard deviation are used to determine the number of data values that fall within a specified interval in a distribution. For example, Chebyshev's theorem (explained later) shows that, for any distribution, at least 75% of the data values will fall within 2 standard deviations of the mean.
4. Finally, the variance and standard deviation are used quite often in inferential statistics. These uses will be shown in later chapters of this textbook.

Coefficient of Variation

Whenever two samples have the same units of measure, the variance and standard deviation for each can be compared directly. For example, suppose an automobile dealer wanted to compare the standard deviation of miles driven for the cars she received as trade-ins on new cars. She found that for a specific year, the standard deviation for Buicks was 422 miles and the standard deviation for Cadillacs was 350 miles. She could say that the variation in mileage was greater in the Buicks. But what if a manager wanted to compare the standard deviations of two different variables, such as the number of sales per salesperson over a 3-month period and the commissions made by these salespeople?

A statistic that allows one to compare standard deviations when the units are different, as in this example, is called the *coefficient of variation.*

The **coefficient of variation,** denoted by CVar, is the standard deviation divided by the mean. The result is expressed as a percentage.

For samples,

$$\text{CVar} = \frac{s}{\overline{X}} \cdot 100\%$$

For populations,

$$\text{CVar} = \frac{\sigma}{\mu} \cdot 100\%$$

Example 3–25

The mean of the number of sales of cars over a 3-month period is 87, and the standard deviation is 5. The mean of the commissions is $5225, and the standard deviation is $773. Compare the variations of the two.

Solution

The coefficients of variation are

$$\text{CVar} = \frac{s}{\overline{X}} = \frac{5}{87} \cdot 100\% = 5.7\% \qquad \text{sales}$$

$$\text{CVar} = \frac{773}{5225} \cdot 100\% = 14.8\% \qquad \text{commissions}$$

Since the coefficient of variation is larger for commissions, the commissions are more variable than the sales.

Example 3–26

The mean for the number of pages of a sample of women’s fitness magazines is 132, with a variance of 23; the mean for the number of advertisements of a sample of women’s fitness magazines is 182, with a variance of 62. Compare the variations.

Solution

The coefficients of variation are

$$\text{CVar} = \frac{\sqrt{23}}{132} \cdot 100\% = 3.6\% \qquad \text{pages}$$

$$\text{CVar} = \frac{\sqrt{62}}{182} \cdot 100\% = 4.3\% \qquad \text{advertisements}$$

The number of advertisements is more variable than the number of pages since the coefficient of variation is larger for advertisements.

Range Rule of Thumb

The range can be used to approximate the standard deviation. The approximation is called the **range rule of thumb.**

The Range Rule of Thumb

A rough estimate of the standard deviation is

$$s \approx \frac{\text{range}}{4}$$

In other words, if the range is divided by 4, an approximate value for the standard deviation is obtained. For example, the standard deviation for the data set 5, 8, 8, 9, 10, 12, and 13 is 2.7, and the range is $13 - 5 = 8$. The range rule of thumb is $s \approx 2$. The range rule of thumb in this case underestimates the standard deviation somewhat; however, it is in the ballpark.

A note of caution should be mentioned here. The range rule of thumb is only an *approximation* and should be used when the distribution of data values is unimodal and roughly symmetric.

The range rule of thumb can be used to estimate the largest and smallest data values of a data set. The smallest data value will be approximately 2 standard deviations below the mean, and the largest data value will be approximately 2 standard deviations above the mean of the data set. The mean for the previous data set is 9.3; hence,

$$\text{Smallest data value} = \overline{X} - 2s = 9.3 - 2(2.8) = 3.7$$

$$\text{Largest data value} = \overline{X} + 2s = 9.3 + 2(2.8) = 14.9$$

Notice that the smallest data value was 5, and the largest data value was 13. Again, these are rough approximations. For many data sets, almost all data values will fall within 2 standard deviations of the mean. Better approximations can be obtained by using Chebyshev's theorem and the empirical rule. These are explained next.

Chebyshev's Theorem

As stated previously, the variance and standard deviation of a variable can be used to determine the spread, or dispersion, of a variable. That is, the larger the variance or standard deviation, the more the data values are dispersed. For example, if two variables measured in the same units have the same mean, say, 70, and variable 1 has a standard deviation of 1.5 while variable 2 has a standard deviation of 10, then the data for variable 2 will be more spread out than the data for variable 1. *Chebyshev's theorem,* developed by the Russian mathematician Chebyshev (1821–1894), specifies the proportions of the spread in terms of the standard deviation.

Chebyshev's theorem The proportion of values from a data set that will fall within k standard deviations of the mean will be at least $1 - 1/k^2$, where k is a number greater than 1 (k is not necessarily an integer).

This theorem states that at least three-fourths, or 75%, of the data values will fall within 2 standard deviations of the mean of the data set. This result is found by substituting $k = 2$ in the expression.

$$1 - \frac{1}{k^2} \qquad \text{or} \qquad 1 - \frac{1}{2^2} = 1 - \frac{1}{4} = \frac{3}{4} = 75\%$$

For the example in which variable 1 has a mean of 70 and a standard deviation of 1.5, at least three-fourths, or 75%, of the data values fall between 67 and 73. These values are found by adding 2 standard deviations to the mean and subtracting 2 standard deviations from the mean, as shown:

$$70 + 2(1.5) = 70 + 3 = 73$$

and

$$70 - 2(1.5) = 70 - 3 = 67$$

For variable 2, at least three-fourths, or 75%, of the data values fall between 50 and 90. Again, these values are found by adding and subtracting, respectively, 2 standard deviations to and from the mean.

$$70 + 2(10) = 70 + 20 = 90$$

and

$$70 - 2(10) = 70 - 20 = 50$$

Furthermore, the theorem states that at least eight-ninths, or 88.89%, of the data values will fall within 3 standard deviations of the mean. This result is found by letting $k = 3$ and substituting in the expression.

$$1 - \frac{1}{k^2} \qquad \text{or} \qquad 1 - \frac{1}{3^2} = 1 - \frac{1}{9} = \frac{8}{9} = 88.89\%$$

For variable 1, at least eight-ninths, or 88.89%, of the data values fall between 65.5 and 74.5, since

$$70 + 3(1.5) = 70 + 4.5 = 74.5$$

and

$$70 - 3(1.5) = 70 - 4.5 = 65.5$$

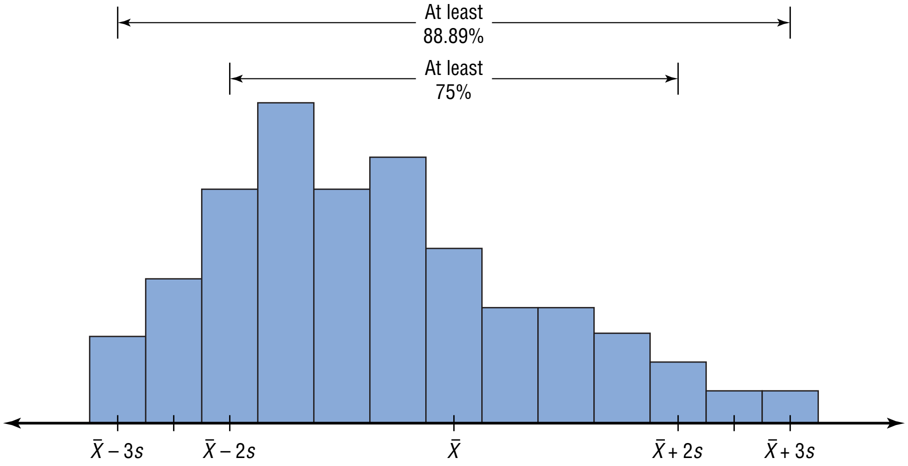
For variable 2, at least eight-ninths, or 88.89%, of the data values fall between 40 and 100.

This theorem can be applied to any distribution regardless of its shape (see Figure 3–3).

Examples 3–27 and 3–28 illustrate the application of Chebyshev's theorem.

Figure 3–3

Chebyshev's Theorem



Example 3–27

The mean price of houses in a certain neighborhood is \$50,000, and the standard deviation is \$10,000. Find the price range for which at least 75% of the houses will sell.

Solution

Chebyshev's theorem states that three-fourths, or 75%, of the data values will fall within 2 standard deviations of the mean. Thus,

$$\$50{,}000 + 2(\$10{,}000) = \$50{,}000 + \$20{,}000 = \$70{,}000$$

and

$$\$50{,}000 - 2(\$10{,}000) = \$50{,}000 - \$20{,}000 = \$30{,}000$$

Hence, at least 75% of all homes sold in the area will have a price range from \$30,000 to \$70,000.

Chebyshev's theorem can be used to find the minimum percentage of data values that will fall between any two given values. The procedure is shown in Example 3–28.

Example 3–28

A survey of local companies found that the mean amount of travel allowance for executives was \$0.25 per mile. The standard deviation was \$0.02. Using Chebyshev's theorem, find the minimum percentage of the data values that will fall between \$0.20 and \$0.30.

Solution

Step 1 Subtract the mean from the larger value.

$$\$0.30 - \$0.25 = \$0.05$$

Step 2 Divide the difference by the standard deviation to get k.

$$k = \frac{0.05}{0.02} = 2.5$$

Step 3 Use Chebyshev's theorem to find the percentage.

$$1 - \frac{1}{k^2} = 1 - \frac{1}{2.5^2} = 1 - \frac{1}{6.25} = 1 - 0.16 = 0.84 \qquad \text{or} \qquad 84\%$$

Hence, at least 84% of the data values will fall between \$0.20 and \$0.30.

The Empirical (Normal) Rule

Chebyshev's theorem applies to any distribution regardless of its shape. However, when a distribution is *bell-shaped* (or what is called *normal*), the following statements, which make up the **empirical rule,** are true.

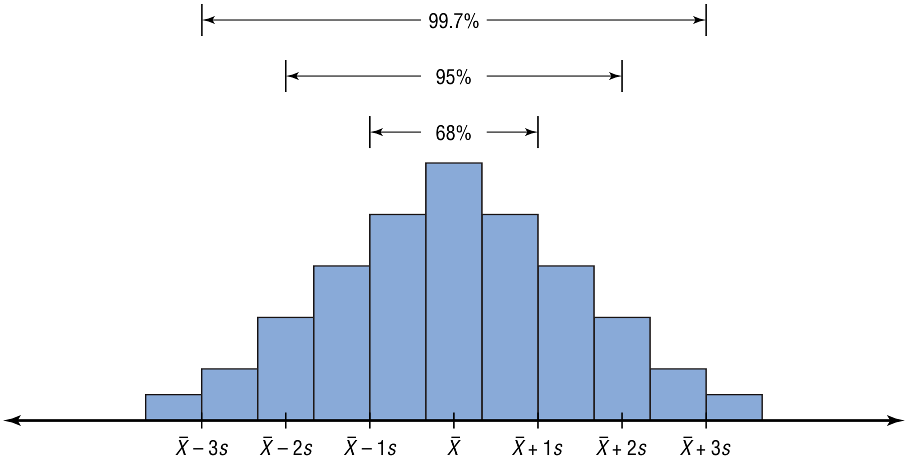
Approximately 68% of the data values will fall within 1 standard deviation of the mean.

Approximately 95% of the data values will fall within 2 standard deviations of the mean.

Approximately 99.7% of the data values will fall within 3 standard deviations of the mean.

Figure 3–4

The Empirical Rule



For example, suppose that the scores on a national achievement exam have a mean of 480 and a standard deviation of 90. If these scores are normally distributed, then approximately 68% will fall between 390 and 570 ($480 + 90 = 570$ and $480 - 90 = 390$). Approximately 95% of the scores will fall between 300 and 660 ($480 + 2 \cdot 90 = 660$ and $480 - 2 \cdot 90 = 300$). Approximately 99.7% will fall between 210 and 750 ($480 + 3 \cdot 90 = 750$ and $480 - 3 \cdot 90 = 210$). See Figure 3–4. (The empirical rule is explained in greater detail in Chapter 7.)

Applying the Concepts 3–3

Blood Pressure

The table lists means and standard deviations. The mean is the number before the plus/minus, and the standard deviation is the number after the plus/minus. The results are from a study attempting to find the average blood pressure of older adults. Use the results to answer the questions.

	Normotensive		**Hypertensive**	
	Men ($n = 1200$)	**Women ($n = 1400$)**	**Men ($n = 1100$)**	**Women ($n = 1300$)**
Age	55 ± 10	55 ± 10	60 ± 10	64 ± 10
Blood pressure (mm Hg)				
Systolic	123 ± 9	121 ± 11	153 ± 17	156 ± 20
Diastolic	78 ± 7	76 ± 7	91 ± 10	88 ± 10

1. Apply Chebyshev's theorem to the systolic blood pressure of normotensive men. At least how many of the men in the study fall within 1 standard deviation of the mean?
2. At least how many of those men in the study fall within 2 standard deviations of the mean?

Assume that blood pressure is normally distributed among older adults. Answer the following questions, using the empirical rule instead of Chebyshev's theorem.

3. Give ranges for the diastolic blood pressure (normotensive and hypertensive) of older women.
4. Do the normotensive, male, systolic blood pressure ranges overlap with the hypertensive, male, systolic, blood pressure ranges?

See page 170 for the answers.

Exercises 3–3

1. What is the relationship between the variance and the standard deviation?

2. Why might the range *not* be the best estimate of variability?

3. What are the symbols used to represent the population variance and standard deviation?

4. What are the symbols used to represent the sample variance and standard deviation?

5. Why is the unbiased estimator of variance used?

6. The three data sets have the same mean and range, but is the variation the same? Prove your answer by computing the standard deviation. Assume the data were obtained from samples.

a. 5, 7, 9, 11, 13, 15, 17
b. 5, 6, 7, 11, 15, 16, 17
c. 5, 5, 5, 11, 17, 17, 17

For Exercises 7–13, find the range, variance, and standard deviation. Assume the data represent samples, and use the shortcut formula for the unbiased estimator to compute the variance and standard deviation.

7. The number of incidents where police were needed for a sample of 10 schools in Allegheny County is 7, 37, 3, 8, 48, 11, 6, 0, 10, 3. Are the data consistent or do they vary? Explain your answer.

Source: U.S. Department of Education.

8. The increases (in cents) in cigarette taxes for 17 states in a 6-month period are

60, 20, 40, 40, 45, 12, 34, 51, 30, 70, 42, 31, 69, 32, 8, 18, 50

Use the range rule of thumb to estimate the standard deviation. Compare the estimate to the actual standard deviation.

Source: Federation of Tax Administrators.

9. The normal daily high temperatures (in degrees Fahrenheit) in January for 10 selected cities are as follows.

50, 37, 29, 54, 30, 61, 47, 38, 34, 61

The normal monthly precipitation (in inches) for these same 10 cities is listed here.

4.8, 2.6, 1.5, 1.8, 1.8, 3.3, 5.1, 1.1, 1.8, 2.5

Which set is more variable?

Source: *N.Y. Times Almanac.*

10. The total surface area (in square miles) for each of six selected Eastern states is listed here.

28,995	PA	37,534	FL
31,361	NY	27,087	VA
20,966	ME	37,741	GA

The total surface area for each of six selected Western states is listed (in square miles).

72,964	AZ	70,763	NV
101,510	CA	62,161	OR
66,625	CO	54,339	UT

Which set is more variable?

Source: *N.Y. Times Almanac.*

11. Shown here are the numbers of stories in the 11 tallest buildings in St. Paul, Minnesota.

32, 36, 46, 20, 32, 18, 16, 34, 26, 27, 26

Shown here are the numbers of stories in the 11 tallest buildings in Chicago, Illinois.

100, 100, 83, 60, 64, 65, 66, 74, 60, 67, 57

Which data set is more variable?

Source: *The World Almanac and Book of Facts.*

12. The following data are the prices of 1 gallon of premium gasoline in U.S. dollars in seven foreign countries.

3.80, 3.80, 3.20, 3.57, 3.62, 3.74, 3.69

Do you think the standard deviation of these data is representative of the population standard deviation of gasoline prices in all foreign countries? Explain your answer.

Source: *Pittsburgh Post Gazette.*

13. The number of weeks on *The New York Times Best Sellers* list for hardcover fiction is

1, 4, 2, 2, 3, 18, 5, 5, 10, 4, 3, 6, 2, 2, 22

Use the range rule of thumb to estimate the standard deviation. Compare the estimate to the actual standard deviation.

Source: *The New York Times Book Review.*

14. Find the range, variance, and standard deviation for the distances of the home runs for McGwire and Sosa, using the data in Exercise 18 in Section 2–2. Compare the ranges and standard deviations. Decide which is more variable or if the variability is about the same. (Use individual data.)

15. Find the range, variance, and standard deviation for each data set in Exercise 11 of Section 3–2. Based on the results, which data set is more variable?

16. The Federal Highway Administration reported the number of deficient bridges in each state. Find the range, variance, and standard deviation.

15,458	1,055	5,008	3,598	8,984
1,337	4,132	10,618	17,361	6,081
6,482	25,090	12,681	16,286	18,832
12,470	17,842	16,601	4,587	47,196
23,205	25,213	23,017	27,768	2,686
7,768	25,825	4,962	22,704	2,694
4,131	13,144	15,582	7,279	12,613
810	13,350	1,208	22,242	7,477
10,902	2,343	2,333	2,979	6,578
14,318	4,773	6,252	734	13,220

Source: *USA TODAY.*

17. Find the range, variance, and standard deviation for the data in Exercise 17 of Section 2–2.

For Exercises 18 through 27, find the variance and standard deviation.

18. For 108 randomly selected college students, this exam score frequency distribution was obtained.

Class limits	Frequency
90–98	6
99–107	22
108–116	43
117–125	28
126–134	9

19. The costs per load (in cents) of 35 laundry detergents tested by a consumer organization are shown here.

Class limits	Frequency
13–19	2
20–26	7
27–33	12
34–40	5
41–47	6
48–54	1
55–61	0
62–68	2

20. Thirty automobiles were tested for fuel efficiency (in miles per gallon). This frequency distribution was obtained.

Class boundaries	Frequency
7.5–12.5	3
12.5–17.5	5
17.5–22.5	15
22.5–27.5	5
27.5–32.5	2

21. The data show the number of murders in 25 selected cities.

Class limits	Frequency
34–96	13
97–159	2
160–222	0
223–285	5
286–348	1
349–411	1
412–474	0
475–537	1
538–600	2

22. In a study of reaction times to a specific stimulus, a psychologist recorded these data (in seconds).

Class limits	Frequency
2.1–2.7	12
2.8–3.4	13
3.5–4.1	7
4.2–4.8	5
4.9–5.5	2
5.6–6.2	1

23. Eighty randomly selected lightbulbs were tested to determine their lifetimes (in hours). This frequency distribution was obtained.

Class boundaries	Frequency
52.5–63.5	6
63.5–74.5	12
74.5–85.5	25
85.5–96.5	18
96.5–107.5	14
107.5–118.5	5

24. The data represent the murder rate per 100,000 individuals in a sample of selected cities in the United States.

Class	Frequency
5–11	8
12–18	5
19–25	7
26–32	1
33–39	1
40–46	3

Source: FBI and U.S. Census Bureau.

25. Eighty randomly selected batteries were tested to determine their lifetimes (in hours). The following frequency distribution was obtained.

Class boundaries	Frequency
62.5–73.5	5
73.5–84.5	14
84.5–95.5	18
95.5–106.5	25
106.5–117.5	12
117.5–128.5	6

Can it be concluded that the lifetimes of these brands of batteries are consistent?

26. Find the variance and standard deviation for the two distributions in Exercise 8 in Section 2–3 and Exercise 18 in Section 2–3. Compare the variation of the data sets. Decide if one data set is more variable than the other.

27. This frequency distribution represents the data obtained from a sample of word processor repairers. The values are the days between service calls on 80 machines.

Class boundaries	Frequency
25.5–28.5	5
28.5–31.5	9
31.5–34.5	32
34.5–37.5	20
37.5–40.5	12
40.5–43.5	2

28. The average score of the students in one calculus class is 110, with a standard deviation of 5; the average score of students in a statistics class is 106, with a standard deviation of 4. Which class is more variable in terms of scores?

29. The data show the lengths (in feet) of suspension bridges in the eastern part of North America and the western part of North America. Compare the variability of the two samples, using the coefficient of variation.

East:	4260, 3500, 2300, 2150, 2000, 1750
West:	4200, 2800, 2310, 1550, 1500, 1207

Source: *World Almanac and Book of Facts.*

30. The average score on an English final examination was 85, with a standard deviation of 5; the average score on a history final exam was 110, with a standard deviation of 8. Which class was more variable?

31. The average age of the accountants at Three Rivers Corp. is 26 years, with a standard deviation of 6 years; the average salary of the accountants is $31,000, with a standard deviation of $4000. Compare the variations of age and income.

32. Using Chebyshev's theorem, solve these problems for a distribution with a mean of 80 and a standard deviation of 10.

a. At least what percentage of values will fall between 60 and 100?

b. At least what percentage of values will fall between 65 and 95?

33. The mean of a distribution is 20 and the standard deviation is 2. Use Chebyshev's theorem.

a. At least what percentage of the values will fall between 10 and 30?

b. At least what percentage of the values will fall between 12 and 28?

34. In a distribution of 200 values, the mean is 50 and the standard deviation is 5. Use Chebyshev's theorem.

a. At least how many values will fall between 30 and 70?

b. At most how many values will be less than 40 or more than 60?

35. A sample of the hourly wages of employees who work in restaurants in a large city has a mean of $5.02 and a standard deviation of $0.09. Using Chebyshev's theorem, find the range in which at least 75% of the data values will fall.

36. A sample of the labor costs per hour to assemble a certain product has a mean of $2.60 and a standard deviation of $0.15. Using Chebyshev's theorem, find the range in which at least 88.89% of the data will lie.

37. A survey of a number of the leading brands of cereal shows that the mean content of potassium per serving is 95 milligrams, and the standard deviation is 2 milligrams. Find the range in which at least 88.89% of the data will fall. Use Chebyshev's theorem.

38. The average score on a special test of knowledge of wood refinishing has a mean of 53 and a standard deviation of 6. Using Chebyshev's theorem, find the range of values in which at least 75% of the scores will lie.

39. The average of the number of trials it took a sample of mice to learn to traverse a maze was 12. The standard deviation was 3. Using Chebyshev's theorem, find the minimum percentage of data values that will fall in the range of 4 to 20 trials.

40. The average cost of a certain type of grass seed is $4.00 per box. The standard deviation is $0.10. Using Chebyshev's theorem, find the minimum percentage of data values that will fall in the range of $3.82 to $4.18.

41. The average U.S. yearly per capita consumption of citrus fruit is 26.8 pounds. Suppose that the distribution of fruit amounts consumed is bell-shaped with a standard deviation equal to 4.2 pounds. What percentage of Americans would you expect to consume more than 31 pounds of citrus fruit per year?

Source: USDA/Economic Research Service.

42. The average full-time faculty member in a post-secondary degree-granting institution works an average of 53 hours per week.

a. If we assume the standard deviation is 2.8 hours, what percentage of faculty members work more than 58.6 hours a week?

b. If we assume a bell-shaped distribution, what percentage of faculty members work more than 58.6 hours a week?

Source: National Center for Education Statistics.

Extending the Concepts

43. For this data set, find the mean and standard deviation of the variable. The data represent the serum cholesterol levels of 30 individuals. Count the number of data values that fall within 2 standard deviations of the mean. Compare this with the number obtained from Chebyshev's theorem. Comment on the answer.

211	240	255	219	204
200	212	193	187	205
256	203	210	221	249
231	212	236	204	187
201	247	206	187	200
237	227	221	192	196

44. For this data set, find the mean and standard deviation of the variable. The data represent the ages of 30 customers who ordered a product advertised on television. Count the number of data values that fall within 2 standard deviations of the mean. Compare this with the number obtained from Chebyshev's theorem. Comment on the answer.

42	44	62	35	20
30	56	20	23	41
55	22	31	27	66
21	18	24	42	25
32	50	31	26	36
39	40	18	36	22

45. Using Chebyshev's theorem, complete the table to find the minimum percentage of data values that fall within *k* standard deviations of the mean.

k	1.5	2	2.5	3	3.5
Percent					

46. Use this data set: 10, 20, 30, 40, 50.

a. Find the standard deviation.

b. Add 5 to each value, and then find the standard deviation.

c. Subtract 5 from each value and find the standard deviation.

d. Multiply each value by 5 and find the standard deviation.

e. Divide each value by 5 and find the standard deviation.

f. Generalize the results of parts *b* through *e*.

g. Compare these results with those in Exercise 38.

47. The mean deviation is found by using this formula:

$$\text{Mean deviation} = \frac{\Sigma|X - \overline{X}|}{n}$$

where

X = value
$\overline{X}$ = mean
n = number of values
$|\ |$ = absolute value

Find the mean deviation for these data.

5, 9, 10, 11, 11, 12, 15, 18, 20, 22

48. A measure to determine the skewness of a distribution is called the *Pearson coefficient of skewness.* The formula is

$$\text{Skewness} = \frac{3(\overline{X} - \text{MD})}{s}$$

The values of the coefficient usually range from −3 to +3. When the distribution is symmetric, the coefficient is zero; when the distribution is positively skewed, it is positive; and when the distribution is negatively skewed, it is negative.

Using the formula, find the coefficient of skewness for each distribution, and describe the shape of the distribution.

a. Mean = 10, median = 8, standard deviation = 3.

b. Mean = 42, median = 45, standard deviation = 4.

c. Mean = 18.6, median = 18.6, standard deviation = 1.5.

d. Mean = 98, median = 97.6, standard deviation = 4.

49. All values of a data set must be within $s\sqrt{n-1}$ of the mean. If a person collected 25 data values that had a mean of 50 and a standard deviation of 3 and you saw that one data value was 67, what would you conclude?

Technology *Step by Step*

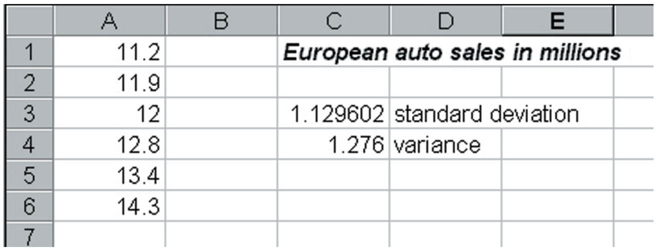
Excel
Step by Step

Finding Measures of Variation

Example XL3–2

To find values that estimate the spread of a distribution of numbers:

1. Enter the numbers in a range (here **A1:A6**). We use the data from Example 3–23 on European automobile sales.
2. For the sample variance, enter **=VAR(A1:A6)** in a blank cell.
3. For the sample standard deviation, enter **=STDEV(A1:A6)** in a blank cell.
4. For the range, you can compute the value **=MAX(A1:A6) − MIN(A1:A6).**



	A	B	C	D	E
1	11.2		*European auto sales in millions*		
2	11.9				
3	12		1.129602	standard deviation	
4	12.8		1.276	variance	
5	13.4				
6	14.3				
7					

There are also functions STDEVP for population standard deviation and VARP for population variances.

3–4 Measures of Position

Objective 3

Identify the position of a data value in a data set, using various measures of position, such as percentiles, deciles, and quartiles.

In addition to measures of central tendency and measures of variation, there are measures of position or location. These measures include standard scores, percentiles, deciles, and quartiles. They are used to locate the relative position of a data value in the data set. For example, if a value is located at the 80th percentile, it means that 80% of the values fall below it in the distribution and 20% of the values fall above it. The *median* is the value that corresponds to the 50th percentile, since one-half of the values fall below it and one-half of the values fall above it. This section discusses these measures of position.

Standard Scores

There is an old saying, "You can't compare apples and oranges." But with the use of statistics, it can be done to some extent. Suppose that a student scored 90 on a music test and 45 on an English exam. Direct comparison of raw scores is impossible, since the exams might not be equivalent in terms of number of questions, value of each question, and so on. However, a comparison of a relative standard similar to both can be made. This comparison uses the mean and standard deviation and is called a standard score or z score. (We also use z scores in later chapters.)

A ***z* score** or **standard score** for a value is obtained by subtracting the mean from the value and dividing the result by the standard deviation. The symbol for a standard score is z. The formula is

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

For samples, the formula is

$$z = \frac{X - \overline{X}}{s}$$

For populations, the formula is

$$z = \frac{X - \mu}{\sigma}$$

The z score represents the number of standard deviations that a data value falls above or below the mean.

For the purpose of this book, it will be assumed that when we find z scores, the data were obtained from samples.

Example 3–29

A student scored 65 on a calculus test that had a mean of 50 and a standard deviation of 10; she scored 30 on a history test with a mean of 25 and a standard deviation of 5. Compare her relative positions on the two tests.

Interesting Fact

The average number of faces that a person learns to recognize and remember during his or her lifetime is 10,000.

Solution

First, find the z scores. For calculus the z score is

$$z = \frac{X - \overline{X}}{s} = \frac{65 - 50}{10} = 1.5$$

For history the z score is

$$z = \frac{30 - 25}{5} = 1.0$$

Since the z score for calculus is larger, her relative position in the calculus class is higher than her relative position in the history class.

Note that if the z score is positive, the score is above the mean. If the z score is 0, the score is the same as the mean. And if the z score is negative, the score is below the mean.

Example 3–30

Find the z score for each test, and state which is higher.

Test A	$X = 38$	$\overline{X} = 40$	$s = 5$
Test B	$X = 94$	$\overline{X} = 100$	$s = 10$

Solution

For test A,

$$z = \frac{X - \overline{X}}{s} = \frac{38 - 40}{5} = -0.4$$

For test B,

$$z = \frac{94 - 100}{10} = -0.6$$

The score for test A is relatively higher than the score for test B.

When all data for a variable are transformed into z scores, the resulting distribution will have a mean of 0 and a standard deviation of 1. A z score, then, is actually the number of standard deviations each value is from the mean for a specific distribution. In Example 3–29, the calculus score of 65 was actually 1.5 standard deviations above the mean of 50. This will be explained in greater detail in Chapter 7.

Percentiles

Percentiles are position measures used in educational and health-related fields to indicate the position of an individual in a group.

Percentiles divide the data set into 100 equal groups.

In many situations, the graphs and tables showing the percentiles for various measures such as test scores, heights, or weights have already been completed. Table 3–3 shows the percentile ranks for scaled scores on the Test of English as a Foreign Language. If a student had a scaled score of 58 for section 1 (listening and comprehension), that student would have a percentile rank of 81. Hence, that student did better than 81% of the students who took section 1 of the exam.

Interesting Facts

The highest recorded temperature on Earth was 136°F in Libya in 1922. The lowest recorded temperature on Earth was −129°F in Antarctica in 1983.

Table 3–3 Percentile Ranks and Scaled Scores on the Test of English as a Foreign Language*

Scaled score	Section 1: Listening comprehension	Section 2: Structure and written expression	Section 3: Vocabulary and reading comprehension	Total scaled score	Percentile rank
68	99	98			
66	98	96	98	660	99
64	96	94	96	640	97
62	92	90	93	620	94
60	87	84	88	600	89
→58	81	76	81	580	82
56	73	68	72	560	73
54	64	58	61	540	62
52	54	48	50	520	50
50	42	38	40	500	39
48	32	29	30	480	29
46	22	21	23	460	20
44	14	15	16	440	13
42	9	10	11	420	9
40	5	7	8	400	5
38	3	4	5	380	3
36	2	3	3	360	1
34	1	2	2	340	1
32		1	1	320	
30		1	1	300	
Mean	51.5	52.2	51.4	517	Mean
S.D.	7.1	7.9	7.5	68	S.D.

*Based on the total group of 1,178,193 examinees tested from July 1989 through June 1991.

Source: Reprinted by permission of Educational Testing Service, the copyright owner.

Figure 3–5

Weights of Girls by Age and Percentile Rankings

Source: Distributed by Mead Johnson Nutritional Division. Reprinted with permission.

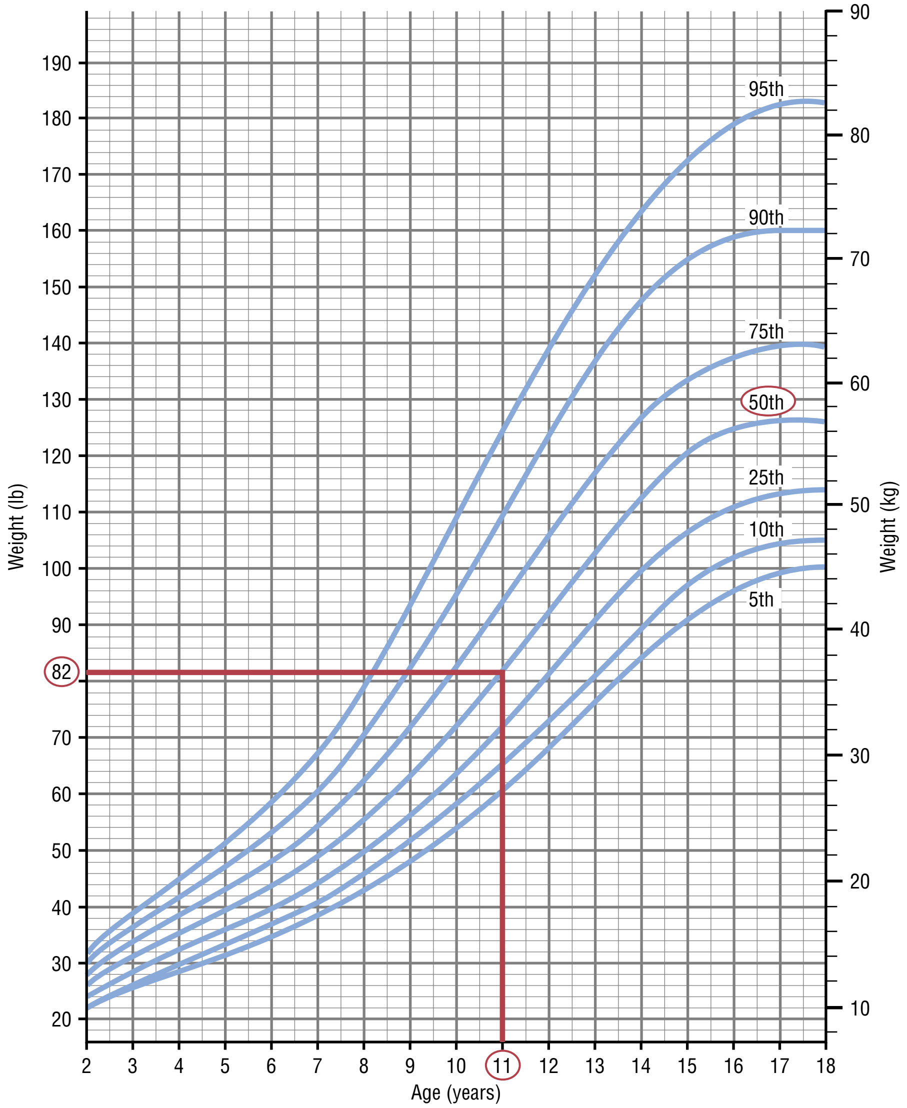


Figure 3–5 shows percentiles in graphical form of weights of girls from ages 2 to 18. To find the percentile rank of an 11-year-old who weighs 82 pounds, start at the 82-pound weight on the left axis and move horizontally to the right. Find 11 on the horizontal axis and move up vertically. The two lines meet at the 50th percentile curved line; hence, an 11-year-old girl who weighs 82 pounds is in the 50th percentile for her age group. If the lines do not meet exactly on one of the curved percentile lines, then the percentile rank must be approximated.

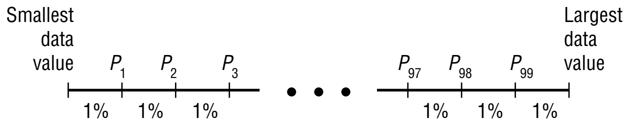
Percentiles are also used to compare an individual's test score with the national norm. For example, tests such as the National Educational Development Test (NEDT) are taken by students in ninth or tenth grade. A student's scores are compared with those of other students locally and nationally by using percentile ranks. A similar test for elementary school students is called the California Achievement Test.

Percentiles are not the same as percentages. That is, if a student gets 72 correct answers out of a possible 100, she obtains a percentage score of 72. There is no indication of her position with respect to the rest of the class. She could have scored the highest, the lowest, or somewhere in between. On the other hand, if a raw score of 72 corresponds to the 64th percentile, then she did better than 64% of the students in her class.

Percentiles are symbolized by

$$P_1, P_2, P_3, \ldots, P_{99}$$

and divide the distribution into 100 groups.



Percentile graphs can be constructed as shown in Example 3–31. Percentile graphs use the same values as the cumulative relative frequency graphs described in Section 2–3, except that the proportions have been converted to percents.

Example 3–31

The frequency distribution for the systolic blood pressure readings (in millimeters of mercury, mm Hg) of 200 randomly selected college students is shown here. Construct a percentile graph.

A Class boundaries	B Frequency	C Cumulative frequency	D Cumulative percent
89.5–104.5	24		
104.5–119.5	62		
119.5–134.5	72		
134.5–149.5	26		
149.5–164.5	12		
164.5–179.5	4		
	200		

Solution

Step 1 Find the cumulative frequencies and place them in column C.

Step 2 Find the cumulative percentages and place them in column D. To do this step, use the formula

$$\text{Cumulative }\% = \frac{\text{cumulative frequency}}{n} \cdot 100\%$$

For the first class,

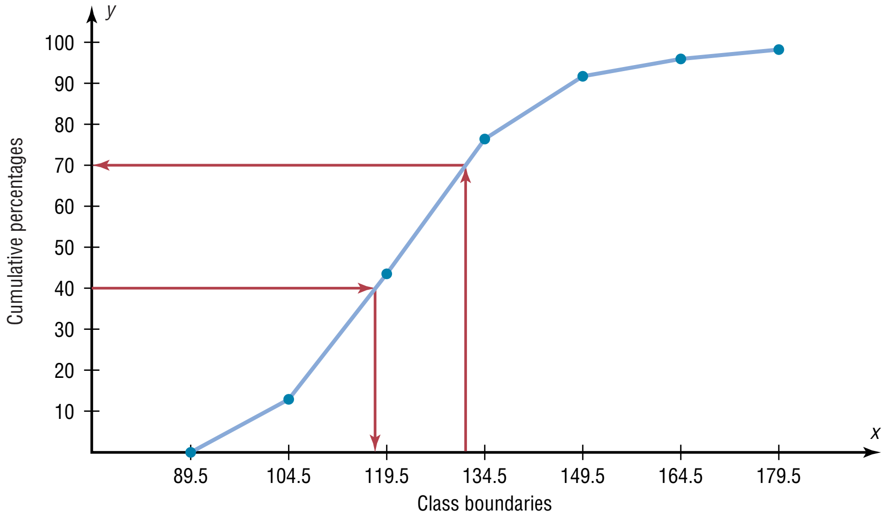
$$\text{Cumulative }\% = \frac{24}{200} \cdot 100\% = 12\%$$

The completed table is shown here.

A Class boundaries	B Frequency	C Cumulative frequency	D Cumulative percent
89.5–104.5	24	24	12
104.5–119.5	62	86	43
119.5–134.5	72	158	79
134.5–149.5	26	184	92
149.5–164.5	12	196	98
164.5–179.5	4	200	100
	200		

Figure 3–6

Percentile Graph for Example 3–31



Step 3 Graph the data, using class boundaries for the *x* axis and the percentages for the *y* axis, as shown in Figure 3–6.

Once a percentile graph has been constructed, one can find the approximate corresponding percentile ranks for given blood pressure values and find approximate blood pressure values for given percentile ranks.

For example, to find the percentile rank of a blood pressure reading of 130, find 130 on the *x* axis of Figure 3–6, and draw a vertical line to the graph. Then move horizontally to the value on the *y* axis. Note that a blood pressure of 130 corresponds to approximately the 70th percentile.

If the value that corresponds to the 40th percentile is desired, start on the *y* axis at 40 and draw a horizontal line to the graph. Then draw a vertical line to the *x* axis and read the value. In Figure 3–6, the 40th percentile corresponds to a value of approximately 118. Thus, if a person has a blood pressure of 118, he or she is at the 40th percentile.

Finding values and the corresponding percentile ranks by using a graph yields only approximate answers. Several mathematical methods exist for computing percentiles for data. These methods can be used to find the approximate percentile rank of a data value or to find a data value corresponding to a given percentile. When the data set is large (100 or more), these methods yield better results. Examples 3–32 through 3–35 show these methods.

Percentile Formula

The percentile corresponding to a given value X is computed by using the following formula:

$$\text{Percentile} = \frac{(\text{number of values below } X) + 0.5}{\text{total number of values}} \cdot 100\%$$

Example 3–32

A teacher gives a 20-point test to 10 students. The scores are shown here. Find the percentile rank of a score of 12.

18, 15, 12, 6, 8, 2, 3, 5, 20, 10

Solution

Arrange the data in order from lowest to highest.

2, 3, 5, 6, 8, 10, 12, 15, 18, 20

Then substitute into the formula.

$$\text{Percentile} = \frac{(\text{number of values below } X) + 0.5}{\text{total number of values}} \cdot 100\%$$

Since there are six values below a score of 12, the solution is

$$\text{Percentile} = \frac{6 + 0.5}{10} \cdot 100\% = 65\text{th percentile}$$

Thus, a student whose score was 12 did better than 65% of the class.

Note: One assumes that a score of 12 in Example 3–32, for instance, means theoretically any value between 11.5 and 12.5.

Example 3–33

Using the data in Example 3–32, find the percentile rank for a score of 6.

Solution

There are three values below 6. Thus

$$\text{Percentile} = \frac{3 + 0.5}{10} \cdot 100\% = 35\text{th percentile}$$

A student who scored 6 did better than 35% of the class.

Examples 3–34 amd 3–35 show a procedure for finding a value corresponding to a given percentile.

Example 3–34

Using the scores in Example 3–32, find the value corresponding to the 25th percentile.

Solution

Step 1 Arrange the data in order from lowest to highest.

2, 3, 5, 6, 8, 10, 12, 15, 18, 20

Step 2 Compute

$$c = \frac{n \cdot p}{100}$$

where
n = total number of values
p = percentile

Thus,

$$c = \frac{10 \cdot 25}{100} = 2.5$$

Step 3 If c is not a whole number, round it up to the next whole number; in this case, $c = 3$. (If c is a whole number, see Example 3–35.) Start at the lowest value and count over to the third value, which is 5. Hence, the value 5 corresponds to the 25th percentile.

Example 3–35

Using the data set in Example 3–32, find the value that corresponds to the 60th percentile.

Solution

Step 1 Arrange the data in order from smallest to largest.

2, 3, 5, 6, 8, 10, 12, 15, 18, 20

Step 2 Substitute in the formula.

$$c = \frac{n \cdot p}{100} = \frac{10 \cdot 60}{100} = 6$$

Step 3 If c is a whole number, use the value halfway between the c and $c + 1$ values when counting up from the lowest value—in this case, the 6th and 7th values.

2, 3, 5, 6, 8, 10, 12, 15, 18, 20
↗ ↖
6th value 7th value

The value halfway between 10 and 12 is 11. Find it by adding the two values and dividing by 2.

$$\frac{10 + 12}{2} = 11$$

Hence, 11 corresponds to the 60th percentile. Anyone scoring 11 would have done better than 60% of the class.

The steps for finding a value corresponding to a given percentile are summarized in this Procedure Table.

Procedure Table

Finding a Data Value Corresponding to a Given Percentile

Step 1 Arrange the data in order from lowest to highest.

Step 2 Substitute into the formula

$$c = \frac{n \cdot p}{100}$$

where

n = total number of values
p = percentile

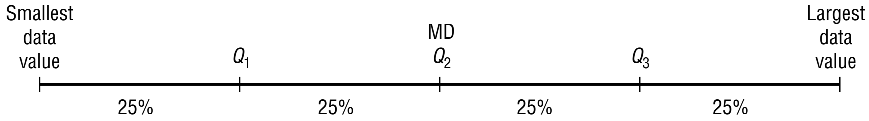
Step 3A If c is not a whole number, round up to the next whole number. Starting at the lowest value, count over to the number that corresponds to the rounded-up value.

Step 3B If c is a whole number, use the value halfway between the cth and $(c + 1)$st values when counting up from the lowest value.

Quartiles and Deciles

Quartiles divide the distribution into four groups, separated by Q_1, Q_2, Q_3.

Note that Q_1 is the same as the 25th percentile; Q_2 is the same as the 50th percentile, or the median; Q_3 corresponds to the 75th percentile, as shown:



Quartiles can be computed by using the formula given for computing percentiles on page 139. For Q_1 use $p = 25$. For Q_2 use $p = 50$. For Q_3 use $p = 75$. However, an easier method for finding quartiles is found in this Procedure Table.

Procedure Table

Finding Data Values Corresponding to Q_1, Q_2, and Q_3

Step 1 Arrange the data in order from lowest to highest.

Step 2 Find the median of the data values. This is the value for Q_2.

Step 3 Find the median of the data values that fall below Q_2. This is the value for Q_1.

Step 4 Find the median of the data values that fall above Q_2. This is the value for Q_3

Example 3–36 shows how to find the values of Q_1, Q_2, and Q_3.

Example 3–36

Find Q_1, Q_2, and Q_3 for the data set 15, 13, 6, 5, 12, 50, 22, 18.

Solution

Step 1 Arrange the data in order.

5, 6, 12, 13, 15, 18, 22, 50

Step 2 Find the median (Q_2).

5, 6, 12, 13, 15, 18, 22, 50
↑
MD

$$\text{MD} = \frac{13 + 15}{2} = 14$$

Step 3 Find the median of the data values less than 14.

5, 6, 12, 13
↑
Q_1

$$Q_1 = \frac{6 + 12}{2} = 9$$

So Q_1 is 9.

Step 4 Find the median of the data values greater than 14.

$$15, 18, \underset{\underset{\textstyle Q_3}{\uparrow}}{\,}22, 50$$

$$Q_3 = \frac{18 + 22}{2} = 20$$

Here Q_3 is 20. Hence, $Q_1 = 9$, $Q_2 = 14$, and $Q_3 = 20$.

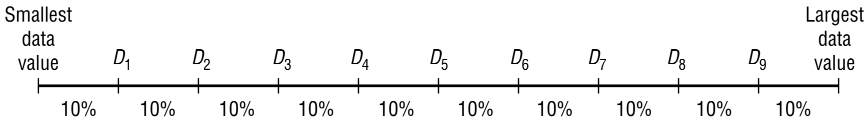
Unusual Stat

Of the alcoholic beverages consumed in the United States, 85% is beer.

In addition to dividing the data set into four groups, quartiles can be used as a rough measurement of variability. The **interquartile range (IQR)** is defined as the difference between Q_1 and Q_3 and is the range of the middle 50% of the data.

The interquartile range is used to identify outliers, and it is also used as a measure of variability in exploratory data analysis, as shown in Section 3–5.

Deciles divide the distribution into 10 groups, as shown. They are denoted by D_1, D_2, etc.



Note that D_1 corresponds to P_{10}; D_2 corresponds to P_{20}; etc. Deciles can be found by using the formulas given for percentiles. Taken altogether then, these are the relationships among percentiles, deciles, and quartiles.

Deciles are denoted by $D_1, D_2, D_3, \ldots, D_9$, and they correspond to $P_{10}, P_{20}, P_{30}, \ldots, P_{90}$.

Quartiles are denoted by Q_1, Q_2, Q_3 and they correspond to P_{25}, P_{50}, P_{75}.

The median is the same as P_{50} or Q_2 or D_5.

The position measures are summarized in Table 3–4.

Table 3–4 Summary of Position Measures

Measure	Definition	Symbol(s)
Standard score or *z* score	Number of standard deviations that a data value is above or below the mean	z
Percentile	Position in hundredths that a data value holds in the distribution	P_n
Decile	Position in tenths that a data value holds in the distribution	D_n
Quartile	Position in fourths that a data value holds in the distribution	Q_n

Outliers

A data set should be checked for extremely high or extremely low values. These values are called *outliers.*

An **outlier** is an extremely high or an extremely low data value when compared with the rest of the data values.

An outlier can strongly affect the mean and standard deviation of a variable. For example, suppose a researcher mistakenly recorded an extremely high data value. This value would then make the mean and standard deviation of the variable much larger than they really were. Outliers can have an effect on other statistics as well.

There are several ways to check a data set for outliers. One method is shown in this Procedure Table.

Procedure Table

Procedure for Identifying Outliers

Step 1 Arrange the data in order and find Q_1 and Q_3.

Step 2 Find the interquartile range: $\text{IQR} = Q_3 - Q_1$.

Step 3 Multiply the IQR by 1.5.

Step 4 Subtract the value obtained in step 3 from Q_1 and add the value to Q_3.

Step 5 Check the data set for any data value that is smaller than $Q_1 - 1.5(\text{IQR})$ or larger than $Q_3 + 1.5(\text{IQR})$.

This procedure is shown in Example 3–37.

Example 3–37

Check the following data set for outliers.

5, 6, 12, 13, 15, 18, 22, 50

Solution

The data value 50 is extremely suspect. These are the steps in checking for an outlier.

Step 1 Find Q_1 and Q_3. This was done in Example 3–36; Q_1 is 9 and Q_3 is 20.

Step 2 Find the interquartile range (IQR), which is $Q_3 - Q_1$.

$$\text{IQR} = Q_3 - Q_1 = 20 - 9 = 11$$

Step 3 Multiply this value by 1.5.

$$1.5(11) = 16.5$$

Step 4 Subtract the value obtained in step 3 from Q_1, and add the value obtained in step 3 to Q_3.

$$9 - 16.5 = -7.5 \qquad \text{and} \qquad 20 + 16.5 = 36.5$$

Step 5 Check the data set for any data values that fall outside the interval from -7.5 to 36.5. The value 50 is outside this interval; hence, it can be considered an outlier.

There are several reasons why outliers may occur. First, the data value may have resulted from a measurement or observational error. Perhaps the researcher measured the variable incorrectly. Second, the data value may have resulted from a recording error. That is, it may have been written or typed incorrectly. Third, the data value may have been obtained from a subject that is not in the defined population. For example, suppose test scores were obtained from a seventh-grade class, but a student in that class was actually in the sixth grade and had special permission to attend the class. This student might have scored extremely low on that particular exam on that day. Fourth, the data value might be a legitimate value that occurred by chance (although the probability is extremely small).

There are no hard-and-fast rules on what to do with outliers, nor is there complete agreement among statisticians on ways to identify them. Obviously, if they occurred as a result of an error, an attempt should be made to correct the error or else the data value should be omitted entirely. When they occur naturally by chance, the statistician must make a decision about whether to include them in the data set.

When a distribution is normal or bell-shaped, data values that are beyond 3 standard deviations of the mean can be considered suspected outliers.

Applying the Concepts 3–4

Determining Dosages

In an attempt to determine necessary dosages of a new drug (HDL) used to control sepsis, assume you administer varying amounts of HDL to 40 mice. You create four groups and label them *low dosage, moderate dosage, large dosage,* and *very large dosage*. The dosages also vary within each group. After the mice are injected with the HDL and the sepsis bacteria, the time until the onset of sepsis is recorded. Your job as a statistician is to effectively communicate the results of the study.

1. Which measures of position could be used to help describe the data results?
2. If 40% of the rats in the top quartile survived after the injection, how many mice would that be?
3. What information can be given from using percentiles?
4. What information can be given from using quartiles?
5. What information can be given from using standard scores?

See page 170 for the answers.

Exercises 3–4

1. What is a z score?

2. Define *percentile rank.*

3. What is the difference between a percentage and a percentile?

4. Define *quartile.*

5. What is the relationship between quartiles and percentiles?

6. What is a decile?

7. How are deciles related to percentiles?

8. To which percentile, quartile, and decile does the median correspond?

9. If the mean value of major league teams is $127 million and the standard deviation is $9 million, find the corresponding z score for each team's value.

a. 136
b. 109
c. 104.5
d. 113.5
e. 133

10. The reaction time to a stimulus for a certain test has a mean of 2.5 seconds and a standard deviation of 0.3 second. Find the corresponding z score for each reaction time.

a. 2.7
b. 3.9
c. 2.8
d. 3.1
e. 2.2

11. A final examination for a psychology course has a mean of 84 and a standard deviation of 4. Find the corresponding z score for each raw score.

a. 87
b. 79
c. 93
d. 76
e. 82

12. An aptitude test has a mean of 220 and a standard deviation of 10. Find the corresponding z score for each exam score.

a. 200
b. 232
c. 218
d. 212
e. 225

13. Which of the following exam scores has a better relative position?

a. A score of 42 on an exam with $\overline{X} = 39$ and $s = 4$.
b. A score of 76 on an exam with $\overline{X} = 71$ and $s = 3$.

14. A student scores 60 on a mathematics test that has a mean of 54 and a standard deviation of 3, and she scores 80 on a history test with a mean of 75 and a standard deviation of 2. On which test did she perform better?

15. Which score indicates the highest relative position?

a. A score of 3.2 on a test with $\overline{X} = 4.6$ and $s = 1.5$.
b. A score of 630 on a test with $\overline{X} = 800$ and $s = 200$.
c. A score of 43 on a test with $\overline{X} = 50$ and $s = 5$.

16. This distribution represents the data for weights of fifth-grade boys. Find the approximate weights corresponding to each percentile given by constructing a percentile graph.

Weight (pounds)	Frequency
52.5–55.5	9
55.5–58.5	12
58.5–61.5	17
61.5–64.5	22
64.5–67.5	15

a. 25th
b. 60th
c. 80th
d. 95th

17. For the data in Exercise 16, find the approximate percentile ranks of the following weights.

a. 57 pounds
b. 62 pounds
c. 64 pounds
d. 59 pounds

18. **(ans)** The data shown represent the scores on a national achievement test for a group of tenth-grade students. Find the approximate percentile ranks of these scores by constructing a percentile graph.

a. 220
b. 245
c. 276
d. 280
e. 300

Score	Frequency
196.5–217.5	5
217.5–238.5	17
238.5–259.5	22
259.5–280.5	48
280.5–301.5	22
301.5–322.5	6

19. For the data in Exercise 18, find the approximate scores that correspond to these percentiles.

a. 15th
b. 29th
c. 43rd
d. 65th
e. 80th

20. **(ans)** The airborne speeds in miles per hour of 21 planes are shown. Find the approximate values that correspond to the given percentiles by constructing a percentile graph.

Class	Frequency
366–386	4
387–407	2
408–428	3
429–449	2
450–470	1
471–491	2
492–512	3
513–533	4
	21

a. 9th
b. 20th
c. 45th
d. 60th
e. 75th

Source: *The World Almanac and Book of Facts.*

21. Using the data in Exercise 20, find the approximate percentile ranks of the following miles per hour (mph).

a. 380 mph
b. 425 mph
c. 455 mph
d. 505 mph
e. 525 mph

22. Find the percentile ranks of each weight in the data set. The weights are in pounds.

78, 82, 86, 88, 92, 97

23. In Exercise 22, what value corresponds to the 30th percentile?

24. Find the percentile rank for each test score in the data set.

12, 28, 35, 42, 47, 49, 50

25. In Exercise 24, what value corresponds to the 60th percentile?

26. Find the percentile rank for each value in the data set. The data represent the values in billions of dollars of the damage of 10 hurricanes.

1.1, 1.7, 1.9, 2.1, 2.2, 2.5, 3.3, 6.2, 6.8, 20.3

Source: Insurance Services Office.

27. What value in Exercise 26 corresponds to the 40th percentile?

28. Find the percentile rank for each test score in the data set.

5, 12, 15, 16, 20, 21

29. What test score in Exercise 28 corresponds to the 33rd percentile?

30. Using the procedure shown in Example 3–37, check each data set for outliers.

a. 16, 18, 22, 19, 3, 21, 17, 20
b. 24, 32, 54, 31, 16, 18, 19, 14, 17, 20
c. 321, 343, 350, 327, 200
d. 88, 72, 97, 84, 86, 85, 100
e. 145, 119, 122, 118, 125, 116
f. 14, 16, 27, 18, 13, 19, 36, 15, 20

31. Another measure of average is called the *midquartile;* it is the numerical value halfway between Q_1 and Q_3, and the formula is

$$\text{Midquartile} = \frac{Q_1 + Q_3}{2}$$

Using this formula and other formulas, find Q_1, Q_2, Q_3, the midquartile, and the interquartile range for each data set.

a. 5, 12, 16, 25, 32, 38
b. 53, 62, 78, 94, 96, 99, 103

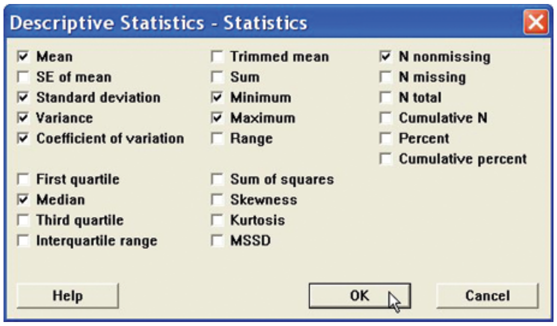
Technology *Step by Step*

MINITAB Step by Step

Calculate Descriptive Statistics from Data

Example MT3–1

1. Enter the data from Example 3–23 into C1 of MINITAB. Name the column **AutoSales.**
2. Select **Stat>Basic Statistics>Display Descriptive Statistics.**
3. The cursor will be blinking in the Variables text box. Double-click C1 AutoSales.
4. Click [Statistics] to view the statistics that can be calculated with this command.
   a) Check the boxes for Mean, Standard deviation, Variance, Coefficient of variation, Median, Minimum, Maximum, and N nonmissing.



Descriptive Statistics - Statistics

☑ Mean
□ SE of mean
☑ Standard deviation
☑ Variance
☑ Coefficient of variation

□ First quartile
☑ Median
□ Third quartile
□ Interquartile range

□ Trimmed mean
□ Sum
☑ Minimum
☑ Maximum
□ Range

□ Sum of squares
□ Skewness
□ Kurtosis
□ MSSD

☑ N nonmissing
□ N missing
□ N total
□ Cumulative N
□ Percent
□ Cumulative percent

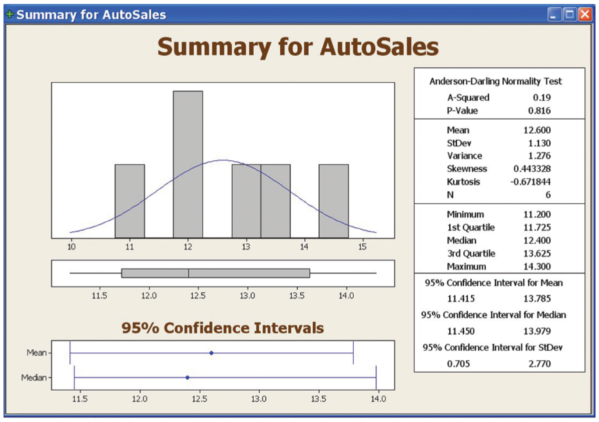
   b) Remove the checks from other options.

5. Click [OK] twice. The results will be displayed in the session window as shown.

```
Descriptive Statistics: AutoSales
Variable   N   Mean  Median    StDev  Variance   CoefVar  Minimum  Maximum
AutoSales  6   12.6    12.4  1.12960     1.276   8.96509     11.2     14.3
```

Session window results are in text format. A high-resolution graphical window displays the descriptive statistics, a histogram, and a boxplot.

6. Select **Stat>Basic Statistics>Graphical Summary.**
7. Double-click C1 AutoSales.
8. Click [OK].



Anderson-Darling Normality Test	
A-Squared	0.19
P-Value	0.816
Mean	12.600
StDev	1.130
Variance	1.276
Skewness	0.443328
Kurtosis	-0.671844
N	6
Minimum	11.200
1st Quartile	11.725
Median	12.400
3rd Quartile	13.625
Maximum	14.300
95% Confidence Interval for Mean	
11.415	13.785
95% Confidence Interval for Median	
11.450	13.979
95% Confidence Interval for StDev	
0.705	2.770

The graphical summary will be displayed in a separate window as shown.

Calculate Descriptive Statistics from a Frequency Distribution

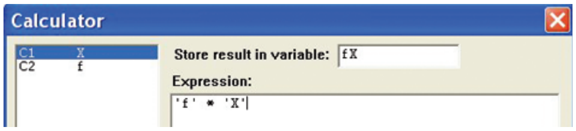
Multiple menu selections must be used to calculate the statistics from a table. We will use data given in Example 3–24.

Enter Midpoints and Frequencies

1. Select **File>New>New Worksheet** to open an empty worksheet.
2. To enter the midpoints into C1, select **Calc>Make Patterned Data>Simple Set of Numbers.**
   a) Type **X** to name the column.
   b) Type in **8** for the First value, **38** for the Last value, and **5** for Steps.
   c) Click [OK].
3. Enter the frequencies in C2. Name the column **f.**

Calculate Columns for f·X and f·X^2

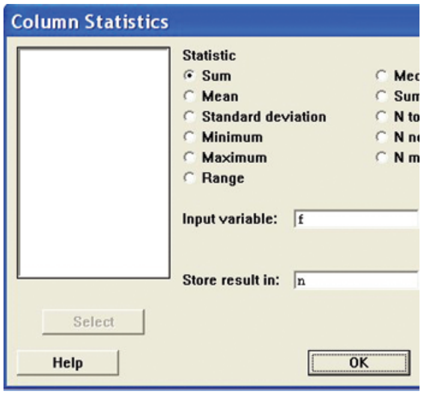
4. Select **Calc>Calculator.**
   a) Type in **fX** for the variable and **f*X** in the Expression dialog box. Click [OK].
   b) Select **Edit>Edit Last Dialog** and type in **fX2** for the variable and **f*X**2** for the expression.
   c) Click [OK]. There are now four columns in the worksheet.


Calculator

C1	X
C2	f

Store result in variable: fX

Expression: 'f' * 'X'


Column Statistics

Statistic

● Sum
○ Mean
○ Standard deviation
○ Minimum
○ Maximum
○ Range
○ Med
○ Sum
○ N to
○ N no
○ N m

Input variable: f

Store result in: n

Calculate the Column Sums

5. Select **Calc>Column Statistics.**

   This command stores results in constants, not columns.

   Click [OK] after each step.

   a) Click the option for Sum; then select C2 f for the Input column, and type **n** for Store result in.

   b) Select **Edit>Edit Last Dialog;** then select C3 fX for the column and type **sumX** for storage.

   c) Edit the last dialog box again. This time select C4 fX2 for the column, then type **sumX2** for storage.

To verify the results, navigate to the Project Manager window, then the constants folder of the worksheet. The sums are 20, 490, and 13,310.

Calculate the Mean, Variance, and Standard Deviation

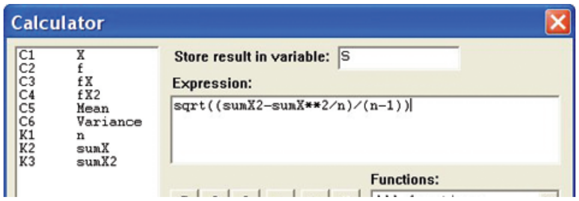
6. Select **Calc>Calculator.**

   a) Type **Mean** for the variable, then click in the box for the Expression and type **sumX/n.** Click [OK]. If you double-click the constants instead of typing them, single quotes will surround the names. The quotes are not required unless the column name has spaces.

   b) Click the **EditLast Dialog** icon and type **Variance** for the variable.

   c) In the expression box type in

   (sumX2-sumX2/n)/(n-1)**


Calculator

C1	X
C2	f
C3	fX
C4	fX2
C5	Mean
C6	Variance
K1	n
K2	sumX
K3	sumX2

Store result in variable: S

Expression: sqrt((sumX2-sumX**2/n)/(n-1))

   d) Edit the last dialog box and type **S** for the variable. In the expression box, drag the mouse over the previous expression to highlight it.

   e) Click the button in the keypad for parentheses. Type **SQRT** at the beginning of the line, upper- or lowercase will work. The expression should be SQRT((sumX2-sumX**2/n)/(n-1)).

   f) Click [OK].

Display Results

   g) Select **Data>Display Data,** then highlight all columns and constants in the list.

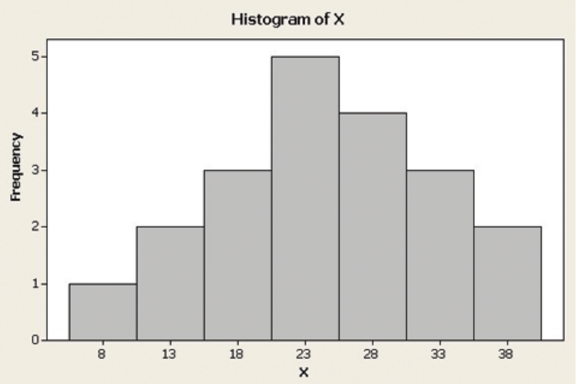
   h) Click [Select] then [OK].

The session window will display all our work! Create the histogram with instructions from Chapter 2.

Data Display

```
n       20.0000
sumX   490.000
sumX2 13310.0

Row   X   f   fX    fX2   Mean   Variance        S
  1   8   1    8     64   24.5    68.6842  8.28759
  2  13   2   26    338
  3  18   3   54    972
  4  23   5  115   2645
  5  28   4  112   3136
  6  33   3   99   3267
  7  38   2   76   2888
```



TI-83 Plus or TI-84 Plus
Step by Step

Calculating Descriptive Statistics

To calculate various descriptive statistics:

1. Enter data into L_1.
2. Press **STAT** to get the menu.
3. Press ▶ to move cursor to CALC; then press **1** for 1-Var Stats.
4. Press **2nd [L_1],** then **ENTER.**

The calculator will display

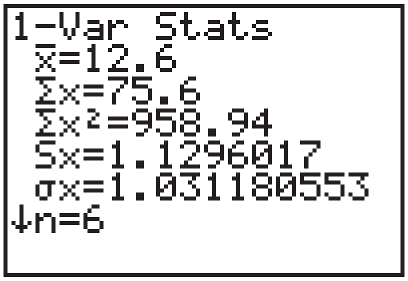
$\bar{x}$	sample mean
Σx	sum of the data values
Σx^2	sum of the squares of the data values
S_x	sample standard deviation
σ_x	population standard deviation
n	number of data values
min$X$	smallest data value
Q_1	lower quartile
Med	median
Q_3	upper quartile
max$X$	largest data value

Example TI3–1

Find the various descriptive statistics for the auto sales data from Example 3–23:

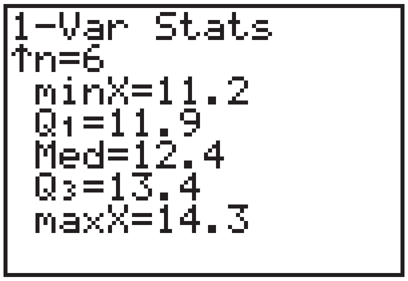
11.2, 11.9, 12.0, 12.8, 13.4, 14.3

Output



```
1-Var Stats
 x̄=12.6
 Σx=75.6
 Σx²=958.94
 Sx=1.1296017
 σx=1.031180553
↓n=6
```

Output



```
1-Var Stats
↑n=6
 minX=11.2
 Q₁=11.9
 Med=12.4
 Q₃=13.4
 maxX=14.3
```

Following the steps just shown, we obtain these results, as shown on the screen:

The mean is 12.6.

The sum of x is 75.6.

The sum of x^2 is 958.94.

The sample standard deviation S_x is 1.1296017.

The population standard deviation σ_x is 1.031180553.

The sample size n is 6.

The smallest data value is 11.2.

Q_1 is 11.9.

The median is 12.4.

Q_3 is 13.4.

The largest data value is 14.3.

To calculate the mean and standard deviation from grouped data:

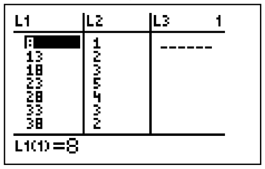
1. Enter the midpoints into L_1.
2. Enter the frequencies into L_2.
3. Press **STAT** to get the menu.
4. Use the arrow keys to move the cursor to CALC; then press **1** for 1-Var Stats.
5. Press **2nd [L1], 2nd [L2],** then **ENTER.**

Example TI3–2

Calculate the mean and standard deviation for the data given in Examples 3–3 and 3–24.

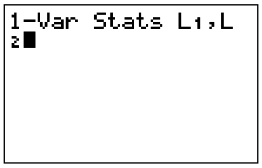
Class	Frequency	Midpoint
5.5–10.5	1	8
10.5–15.5	2	13
15.5–20.5	3	18
20.5–25.5	5	23
25.5–30.5	4	28
30.5–35.5	3	33
35.5–40.5	2	38

Input



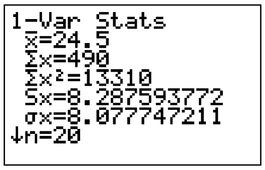
```
L1      |L2     |L3      1
--------+-------+---------
 8      | 1     | ------
 13     | 2     |
 18     | 3     |
 23     | 5     |
 28     | 4     |
 33     | 3     |
 38     | 2     |
L1(1)=8
```

Input



```
1-Var Stats L1,L
2
```

Output

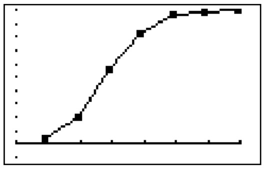


```
1-Var Stats
 x̄=24.5
 Σx=490
 Σx²=13310
 Sx=8.287593772
 σx=8.077747211
↓n=20
```

The sample mean is 24.5, and the sample standard deviation is 8.287593772.

To graph a percentile graph, follow the procedure for an ogive but use the cumulative percent in L_2, 100 for $Y_{\max}$, and the data from Example 3–31.

Output



Excel
Step by Step

Descriptive Statistics in Excel

Example XL3–3

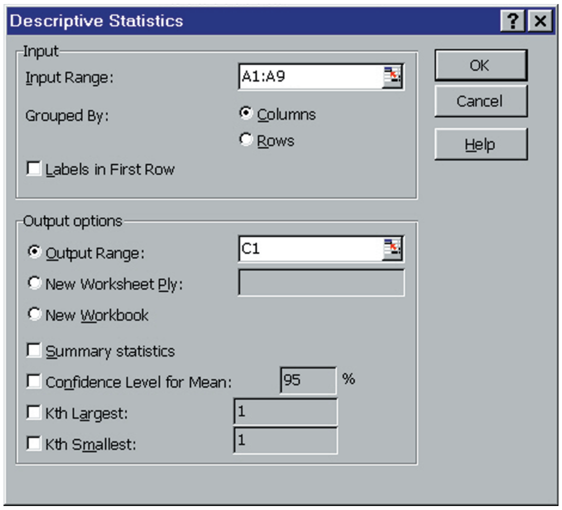
Excel’s Data Analysis options include an item called Descriptive Statistics that reports all the standard measures of a data set.

1. Enter the data set shown (nine numbers) in column A of a new worksheet.

   12 17 15 16 16 14 18 13 10

2. Select **Tools>Data Analysis.**
3. Use these data (A1:A9) as the Input Range in the Descriptive Statistics dialog box.
4. Check the Summary statistics option and click [OK].

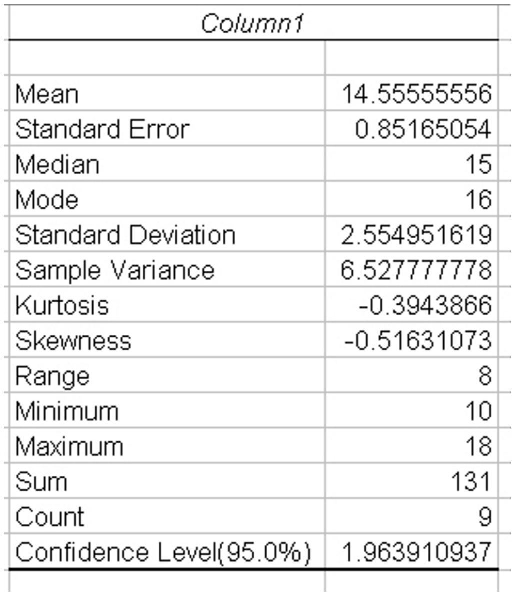
Descriptive Statistics Dialog Box



Input
Input Range: A1:A9
Grouped By: ● Columns ○ Rows
□ Labels in First Row

Output options
● Output Range: C1
○ New Worksheet Ply:
○ New Workbook
□ Summary statistics
□ Confidence Level for Mean: 95 %
□ Kth Largest: 1
□ Kth Smallest: 1

Here’s the summary output for this data set. Note that this one operation reports most of the statistics used in this chapter.



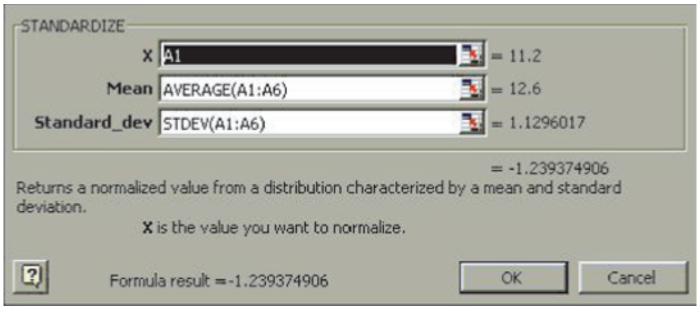
Column1	
Mean	14.55555556
Standard Error	0.85165054
Median	15
Mode	16
Standard Deviation	2.554951619
Sample Variance	6.527777778
Kurtosis	-0.3943866
Skewness	-0.51631073
Range	8
Minimum	10
Maximum	18
Sum	131
Count	9
Confidence Level(95.0%)	1.963910937

Measures of Position

Enter the data from Example 3–23 in column A.

To find the z score for a value in a set of data:

1. Select cell B1 on the worksheet.
2. From the paste function (f_x) icon, select Statistical from the function category. Then select the STANDARDIZE function.
3. Type in **A1** in the X box.
4. Type in **average(A1:A6)** in the Mean box.
5. Type in **stdev(A1:A6)** in the Standard_dev box. Then click [OK].
6. Repeat this procedure for each data value in column A.



STANDARDIZE

X A1 = 11.2

Mean AVERAGE(A1:A6) = 12.6

Standard_dev STDEV(A1:A6) = 1.1296017

= -1.239374906

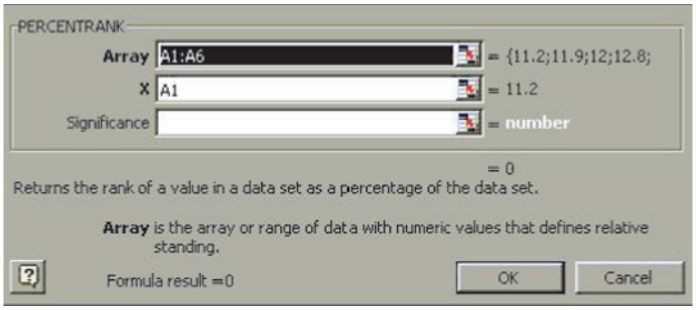
Returns a normalized value from a distribution characterized by a mean and standard deviation.

X is the value you want to normalize.

Formula result =-1.239374906

To find the percentile rank for a value in a set of data:

1. Select cell C1 on the worksheet.
2. From the paste function icon, select Statistical from the function category. Then select the PERCENTRANK function.
3. Type in **A1:A6** in the Array box.
4. Type **A1** in the X box, then click [OK].



PERCENTRANK

Array A1:A6 = {11.2;11.9;12;12.8;

X A1 = 11.2

Significance = number

= 0

Returns the rank of a value in a data set as a percentage of the data set.

Array is the array or range of data with numeric values that defines relative standing.

Formula result =0

3–5

Objective 4

Use the techniques of exploratory data analysis, including boxplots and five-number summaries, to discover various aspects of data.

Exploratory Data Analysis

In traditional statistics, data are organized by using a frequency distribution. From this distribution various graphs such as the histogram, frequency polygon, and ogive can be constructed to determine the shape or nature of the distribution. In addition, various statistics such as the mean and standard deviation can be computed to summarize the data.

The purpose of traditional analysis is to confirm various conjectures about the nature of the data. For example, from a carefully designed study, a researcher might want to know if the proportion of Americans who are exercising today has increased from 10 years ago. This study would contain various assumptions about the population, various definitions such as of exercise, and so on.

In **exploratory data analysis (EDA),** data can be organized using a *stem and leaf plot.* (See Chapter 2.) The measure of central tendency used in EDA is the *median.* The measure of variation used in EDA is the *interquartile range* ($Q_3 - Q_1$). In EDA the data are represented graphically using a **boxplot** (sometimes called a box-and-whisker plot). The purpose of exploratory data analysis is to examine data to find out what information can be discovered about the data such as the center and the spread. Exploratory data analysis was developed by John Tukey and presented in his book *Exploratory Data Analysis* (Addison-Wesley, 1977).

The Five-Number Summary and Boxplots

A **boxplot** can be used to graphically represent the data set. These plots involve five specific values:

1. The lowest value of the data set (i.e., minimum)
2. Q_1
3. The median
4. Q_3
5. The highest value of the data set (i.e., maximum)

These values are called a **five-number summary** of the data set.

A **boxplot** is a graph of a data set obtained by drawing a horizontal line from the minimum data value to Q_1, drawing a horizontal line from Q_3 to the maximum data value, and drawing a box whose vertical sides pass through Q_1 and Q_3 with a vertical line inside the box passing through the median or Q_2.

Example 3–38

A stockbroker recorded the number of clients she saw each day over an 11-day period. The data are shown. Construct a boxplot for the data.

33, 38, 43, 30, 29, 40, 51, 27, 42, 23, 31

Solution

Step 1 Arrange the data in order.

23, 27, 29, 30, 31, 33, 38, 40, 42, 43, 51

Step 2 Find the median.

$$\text{23, 27, 29, 30, 31, } \underset{\substack{\uparrow \\ \text{Median}}}{33}\text{, 38, 40, 42, 43, 51}$$

Step 3 Find Q_1.

$$\text{23, 27, } \underset{\substack{\uparrow \\ 29}}{29}\text{, 30, 31}$$

Step 4 Find Q_3.

$$\text{38, 40, } \underset{\substack{\uparrow \\ 42}}{42}\text{, 43, 51}$$

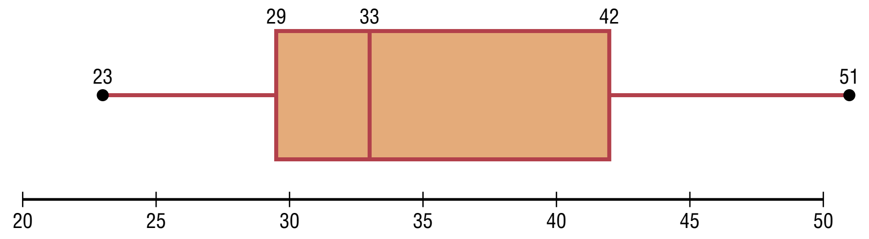


Figure 3–7

Boxplot for Example 3–38

Step 5 Draw a scale for the data on the *x* axis.

Step 6 Locate the lowest value, Q_1, the median, Q_3, and the highest value on the scale.

Step 7 Draw a box around Q_1 and Q_3, draw a vertical line through the median, and connect the upper and lower values, as shown in Figure 3–7.

The box in Figure 3–7 represents the middle 50% of the data, and the lines represent the lower and upper ends of the data.

Information Obtained from a Boxplot

1. *a.* If the median is near the center of the box, the distribution is approximately symmetric.
   b. If the median falls to the left of the center of the box, the distribution is positively skewed.
   c. If the median falls to the right of the center, the distribution is negatively skewed.
2. *a.* If the lines are about the same length, the distribution is approximately symmetric.
   b. If the right line is larger than the left line, the distribution is positively skewed.
   c. If the left line is larger than the right line, the distribution is negatively skewed.

The boxplot in Figure 3–7 indicates that the distribution is slightly positively skewed.

If the boxplots for two or more data sets are graphed on the same axis, the distributions can be compared. To compare the averages, use the location of the medians. To compare the variability, use the interquartile range, i.e., the length of the boxes. Example 3–39 shows this procedure.

Example 3–39

A dietitian is interested in comparing the sodium content of real cheese with the sodium content of a cheese substitute. The data for two random samples are shown. Compare the distributions, using boxplots.

Real cheese				Cheese substitute			
310	420	45	40	270	180	250	290
220	240	180	90	130	260	340	310

Source: The Complete Book of Food Counts.

Solution

Step 1 Find Q_1, MD, and Q_3 for the real cheese data.

40 45 90 180 220 240 310 420

↑ ↑ ↑

Q_1 MD Q_3

$$Q_1 = \frac{45 + 90}{2} = 67.5 \qquad \text{MD} = \frac{180 + 220}{2} = 200$$

$$Q_3 = \frac{240 + 310}{2} = 275$$

Step 2 Find Q_1, MD, and Q_3 for the cheese substitute data.

130 180 250 260 270 290 310 340

↑ ↑ ↑

Q_1 MD Q_3

$$Q_1 = \frac{180 + 250}{2} = 215 \qquad \text{MD} = \frac{260 + 270}{2} = 265$$

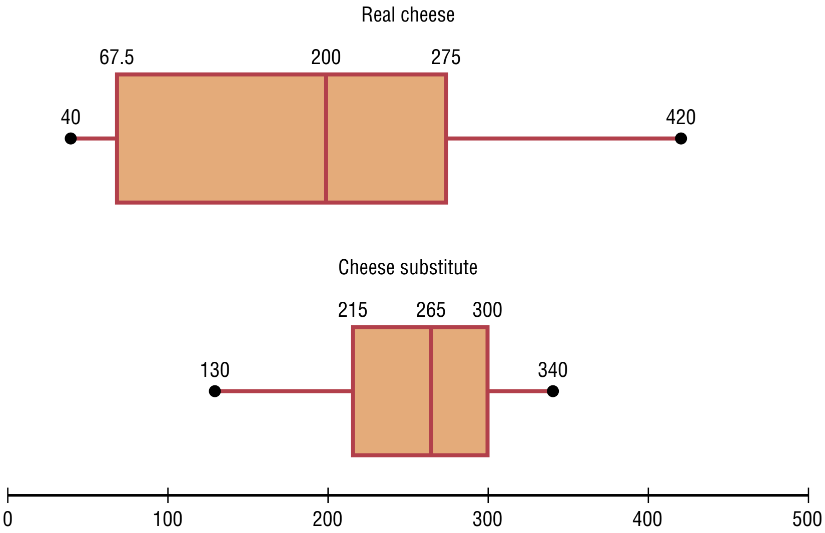
$$Q_3 = \frac{290 + 310}{2} = 300$$

Step 3 Draw the boxplots for each distribution on the same graph. See Figure 3–8.

Step 4 Compare the plots. It is quite apparent that the distribution for the cheese substitute data has a higher median than the median for the distribution for the real cheese data. The variation or spread for the distribution of the real cheese data is larger than the variation for the distribution of the cheese substitute data.

Figure 3–8

Boxplots for Example 3–39



In exploratory data analysis, *hinges* are used instead of quartiles to construct boxplots. When the data set consists of an even number of values, hinges are the same as quartiles. Hinges for a data set with an odd number of values differ somewhat from quartiles. However, since most calculators and computer programs use quartiles, they will be used in this textbook.

Another important point to remember is that the summary statistics (median and interquartile range) used in exploratory data analysis are said to be *resistant statistics.* A **resistant statistic** is relatively less affected by outliers than a *nonresistant statistic.* The mean and standard deviation are nonresistant statistics. Sometimes when a distribution is skewed or contains outliers, the median and interquartile range may more accurately summarize the data than the mean and standard deviation, since the mean and standard deviation are more affected in this case.

Table 3–5 compares the traditional versus the exploratory data analysis approach.

Table 3–5 Traditional versus EDA Techniques

Traditional	Exploratory data analysis
Frequency distribution	Stem and leaf plot
Histogram	Boxplot
Mean	Median
Standard deviation	Interquartile range

Applying the Concepts 3–5

The Noisy Workplace

Assume you work for OSHA (Occupational Safety and Health Administration) and have complaints about noise levels from some of the workers at a state power plant. You charge the power plant with taking decibel readings at six different areas of the plant at different times of the day and week. The results of the data collection are listed. Use boxplots to initially explore the data and make recommendations about which plant areas workers must be provided with protective ear wear. The safe hearing level is at approximately 120 decibels.

Area 1	Area 2	Area 3	Area 4	Area 5	Area 6
30	64	100	25	59	67
12	99	59	15	63	80
35	87	78	30	81	99
65	59	97	20	110	49
24	23	84	61	65	67
59	16	64	56	112	56
68	94	53	34	132	80
57	78	59	22	145	125
100	57	89	24	163	100
61	32	88	21	120	93
32	52	94	32	84	56
45	78	66	52	99	45
92	59	57	14	105	80
56	55	62	10	68	34
44	55	64	33	75	21

See page 170 for the answers.

Exercises 3–5

For Exercises 1–6, identify the five-number summary and find the interquartile range.

1. 8, 12, 32, 6, 27, 19, 54

2. 19, 16, 48, 22, 7

3. 362, 589, 437, 316, 192, 188

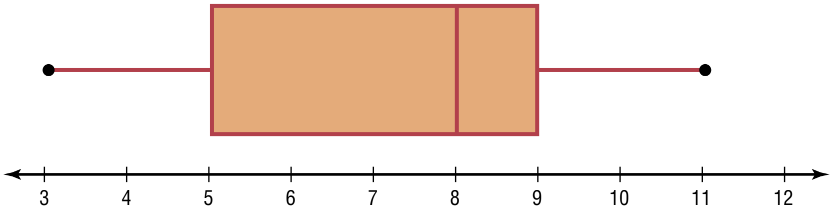
4. 147, 243, 156, 632, 543, 303

5. 14.6, 19.8, 16.3, 15.5, 18.2

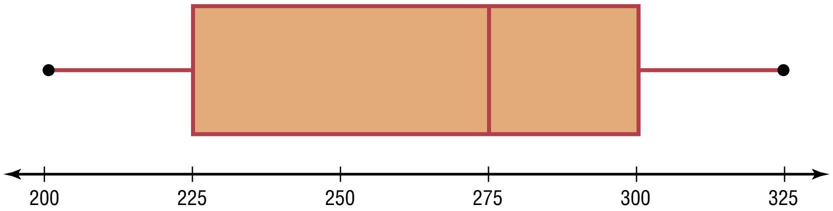
6. 9.7, 4.6, 2.2, 3.7, 6.2, 9.4, 3.8

For Exercises 7–10, use each boxplot to identify the maximum value, minimum value, median, first quartile, third quartile, and interquartile range.

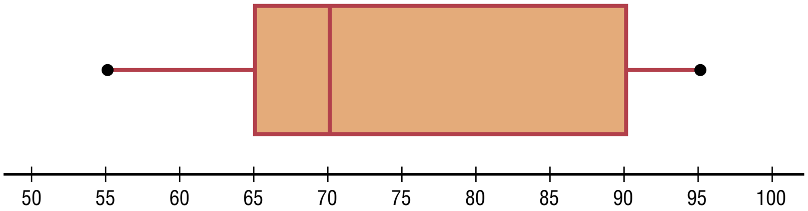
7.



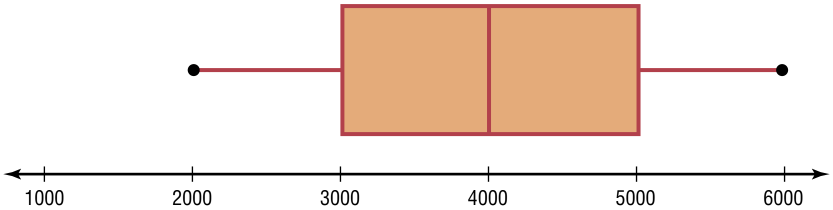
8.



9.



10.



11. Shown next are the sizes of the police forces in the 10 largest cities in the United States in 1993 (the numbers represent hundreds). Construct a boxplot for the data and comment on the shape of the distribution.

29.3, 7.6, 12.1, 4.7, 6.2, 1.9, 3.9, 2.8, 2.0, 1.7

Source: *USA TODAY.*

12. Construct a boxplot for the number of bills enacted by Congress during the last several years. Comment on the shape of the distribution.

88, 245, 153, 241, 170, 410, 136, 241, 198

Source: *USA TODAY.*

13. Construct a boxplot for the following average number of vacation days in selected countries.

42	37	35
34	28	26
13	25	25

Source: World Tourism Organization.

14. Shown here is the number of new theater productions that appeared on Broadway for the past several years. Construct a boxplot for the data and comment on the shape of the distribution.

30	28	33	29	37	39
35	37	37	38	34	

Source: The League of American Theaters and Producers Inc.

15. These data are the number of inches of snow reported in randomly selected U.S. cities for September 1 through January 10. Construct a boxplot and comment on the skewness of the data.

9.8	8.0	13.9	4.4	3.9	21.7
3.2	11.7	24.8	34.1	17.6	15.9

Source: *USA TODAY.*

16. These data represent the volumes in cubic yards of the largest dams in the United States and in South America. Construct a boxplot of the data for each region and compare the distributions.

United States	South America
125,628	311,539
92,000	274,026
78,008	105,944
77,700	102,014
66,500	56,242
62,850	46,563
52,435	
50,000	

Source: *N.Y. Times Almanac.*

17. A 4-month record for the number of tornadoes in 1999–2001 is given here.

a. Which month has the highest mean number of tornadoes for this 3-year period?

b. Which year has the highest mean number of tornadoes for this 4-month period?

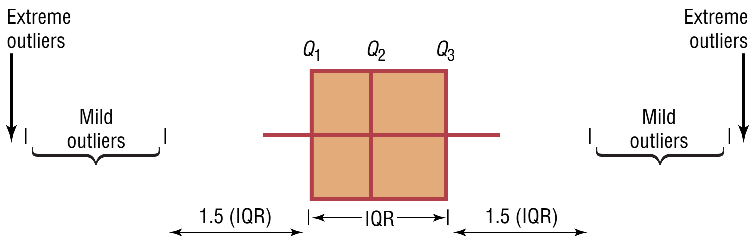
c. Construct three boxplots and compare the distributions.

	2001	2000	1999
April	135	136	177
May	241	241	311
June	248	135	289
July	120	148	102

Source: National Weather Service, Storm Prediction Center.

Extending the Concepts

18. A *modified boxplot* can be drawn by placing a box around Q_1 and Q_3 and then extending the whiskers to the largest and/or smallest values within 1.5 times the interquartile range (i.e., $Q_3 - Q_1$). *Mild outliers* are values between 1.5 (IQR) and 3 (IQR). *Extreme outliers* are data values beyond 3 (IQR).



For the data shown here, draw a modified boxplot and identify any mild or extreme outliers. The data represent the number of unhealthful smog days for a specific year for the highest 10 locations.

97	39	43	66	91
43	54	42	53	39

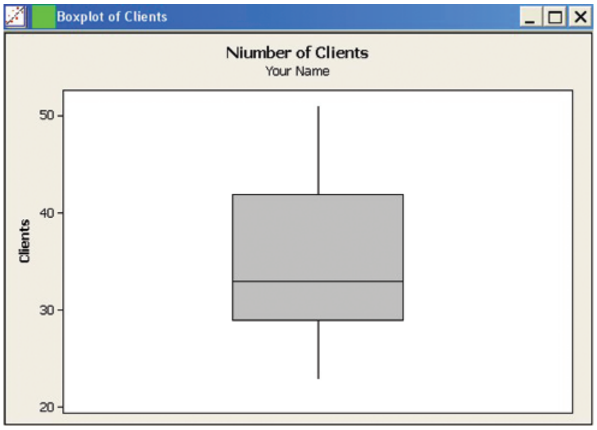
Source: U.S. Public Interest Research Group and Clean Air Network.

Technology *Step by Step*

MINITAB Step by Step

Construct a Boxplot

1. Type in the data for Example 3–38 (the number of clients seen daily by a stockbroker). Label the column **Clients.**
2. Select **Stat>EDA>Boxplot.**
3. Double-click Clients to select it for the Y variable.
4. Click on [Labels].
   a) In the Title 1: of the Title/Footnotes folder, type **Number of Clients.**
   b) Press the [Tab] key and type **Your Name** in the text box for Subtitle 1:.
5. Click [OK] twice. The graph will be displayed in a graph window.

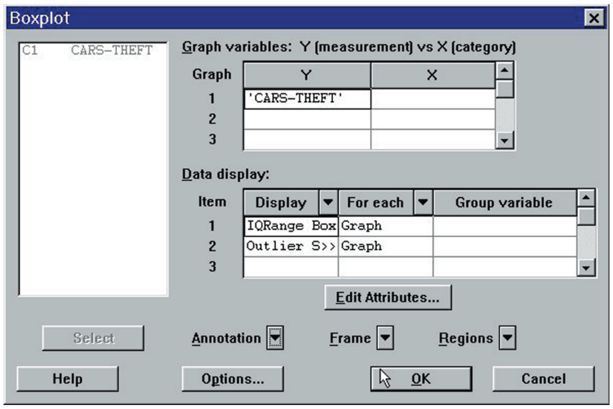


Example MT3–2

1. Enter the data for Example 2–13 in Section 2–4. Label the column **CARS-THEFT.**
2. Select **Stat>EDA>Boxplot.**
3. Double-click CARS-THEFT to select it for the Y variable.
4. Click on the drop-down arrow for Annotation.
5. Click on Title, then enter an appropriate title such as **Car Thefts for Large City, U.S.A.**
6. Click [OK] twice.

A high-resolution graph will be displayed in a graph window.

Boxplot Dialog Box and Boxplot



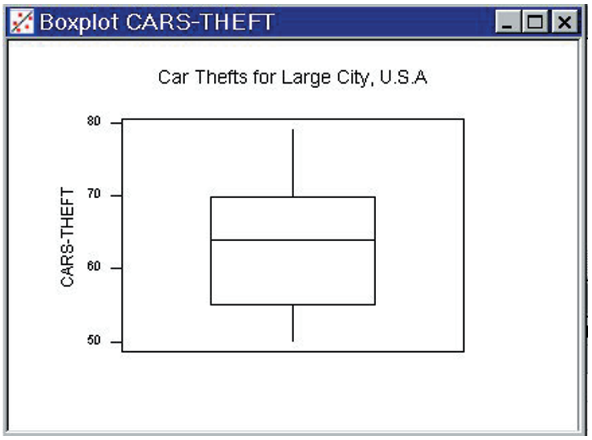
C1 CARS-THEFT

Graph variables: Y (measurement) vs X (category)

Graph	Y	X
1	'CARS-THEFT'	
2		
3		

Data display:

Item	Display	For each	Group variable
1	IQRange Box	Graph	
2	Outlier S>>	Graph	
3			



TI-83 Plus or TI-84 Plus

Step by Step

Constructing a Boxplot

To draw a boxplot:

1. Enter data into L_1.
2. Change values in WINDOW menu, if necessary. (*Note:* Make $X_{\min}$ somewhat smaller than the smallest data value and $X_{\max}$ somewhat larger than the largest data value.) Change $Y_{\min}$ to 0 and $Y_{\max}$ to 1.
3. Press **[2nd] [STAT PLOT]**, then **1** for Plot 1.
4. Press **ENTER** to turn Plot 1 on.
5. Move cursor to Boxplot symbol (fifth graph) on the Type: line, then press **ENTER.**
6. Make sure Xlist is L_1.
7. Make sure Freq is 1.
8. Press **GRAPH** to display the boxplot.
9. Press **TRACE** followed by ◀ or ▶ to obtain the values from the five-number summary on the boxplot.

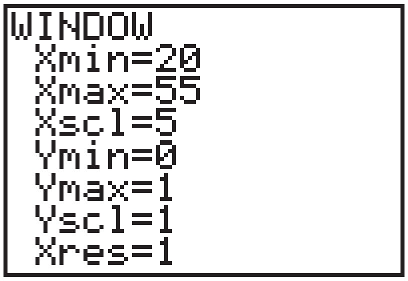
To display two boxplots on the same display, follow the above steps and use the 2: Plot 2 and L_2 symbols.

Example TI3–3

Construct a boxplot for the data values in Example 3–38:

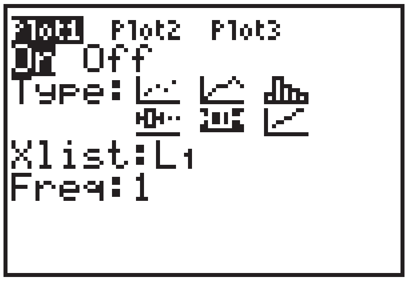
33, 38, 43, 30, 29, 40, 51, 27, 42, 23, 31

Input



```
WINDOW
 Xmin=20
 Xmax=55
 Xscl=5
 Ymin=0
 Ymax=1
 Yscl=1
 Xres=1
```

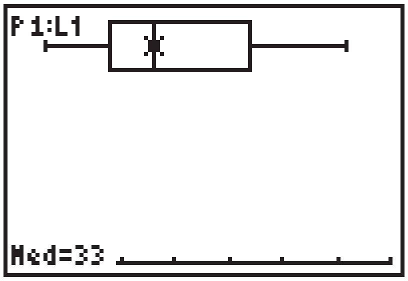
Input



```
Plot1  Plot2  Plot3
On  Off
Type:
Xlist:L1
Freq:1
```

Using the **TRACE** key along with the ◀ and ▶ keys, we obtain the five-number summary. The minimum value is 23; Q_1 is 29; the median is 33; Q_3 is 42; the maximum value is 51.

Output



```
P1:L1
Med=33
```

Excel

Step by Step

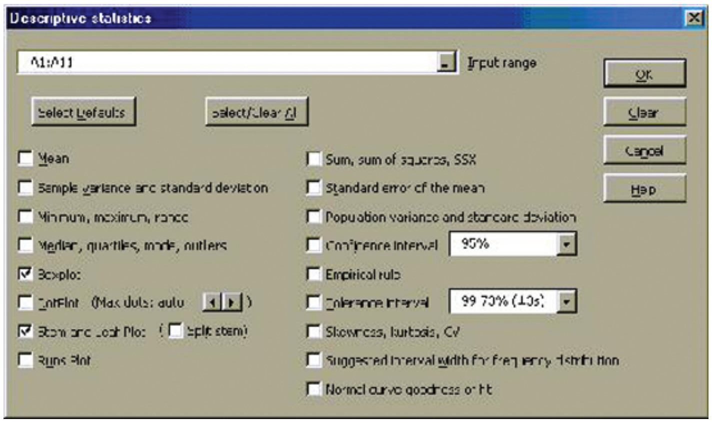
Constructing a Boxplot

Example XL3–4

Excel does not have a procedure to produce stem and leaf plots or boxplots. However, you may construct these plots by using the MegaStat Add-in available on your CD and Online Learning Center. If you have not installed this add-in, do so by following the instructions on page 24.

To obtain a boxplot and stem and leaf plot:

1. Enter the data from Example 3–38 into column A of a new worksheet.
2. Select **MegaStat>Descriptive Statistics.**
3. Enter the cell range **A1:A50** in the Input range.
4. Check the Boxplot and Stem and Leaf plot options. Click [OK].



Descriptive statistics

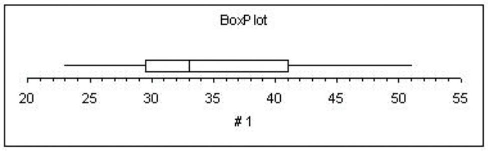
Input range: A1:A11

- □ Mean
- □ Sample variance and standard deviation
- □ Minimum, maximum, range
- □ Median, quartiles, mode, outliers
- ☑ Boxplot
- □ DotPlot (Max dots: auto)
- ☑ Stem and Leaf Plot (□ split stem)
- □ Runs Plot
- □ Sum, sum of squares, SSX
- □ Standard error of the mean
- □ Population variance and standard deviation
- □ Confidence interval 95%
- □ Empirical rule
- □ Tolerance interval 99.73% (±3s)
- □ Skewness, kurtosis, CV
- □ Suggested interval width for frequency distribution
- □ Normal curve goodness of fit

The stem and leaf plot and boxplot that are obtained in the output are featured.

```
Stem and Leaf plot# 1
            stem unit=10
              leaf unit=1

        Frequency   Stem   Leaf
                3      2   3 7 9
                4      3   0 1 3 8
                3      4   0 2 3
                1      5   1
               ──
               11
```



3–6 Summary

This chapter explains the basic ways to summarize data. These include measures of central tendency, measures of variation or dispersion, and measures of position. The three most commonly used measures of central tendency are the mean, median, and mode. The midrange is also used occasionally to represent an average. The three most commonly used measurements of variation are the range, variance, and standard deviation.

The most common measures of position are percentiles, quartiles, and deciles. This chapter explains how data values are distributed according to Chebyshev's theorem and the empirical rule. The coefficient of variation is used to describe the standard deviation in relationship to the mean. These methods are commonly called traditional statistical methods and are primarily used to confirm various conjectures about the nature of the data.

Other methods, such as the boxplot and five-number summaries, are part of exploratory data analysis; they are used to examine data to see what they reveal.

After learning the techniques presented in Chapter 2 and this chapter, you will have a substantial knowledge of descriptive statistics. That is, you will be able to collect, organize, summarize, and present data.

Important Terms

bimodal 103

boxplot 153

Chebyshev's theorem 125

coefficient of variation 124

data array 101

decile 142

empirical rule 127

exploratory data analysis 153

five-number summary 153

interquartile range 142

mean 98

median 101

midrange 106

modal class 104

mode 103

multimodal 103

negatively skewed or left-skewed distribution 109

outlier 142

parameter 98

percentile 135

positively skewed or right-skewed distribution 108

quartile 141

range 116

range rule of thumb 125

resistant statistic 156

standard deviation 118

statistic 98

symmetric distribution 108

unimodal 103

variance 118

weighted mean 107

z score or standard score 133

Important Formulas

Formula for the mean for individual data:

$$\overline{X} = \frac{\Sigma X}{n} \qquad \mu = \frac{\Sigma X}{N}$$

Formula for the mean for grouped data:

$$\overline{X} = \frac{\Sigma f \cdot X_m}{n}$$

Formula for the weighted mean:

$$\overline{X} = \frac{\Sigma wX}{\Sigma w}$$

Formula for the midrange:

$$\text{MR} = \frac{\text{lowest value} + \text{highest value}}{2}$$

Formula for the range:

$$R = \text{highest value} - \text{lowest value}$$

Formula for the variance for population data:

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$$

Formula for the variance for sample data (shortcut formula for the unbiased estimator):

$$s^2 = \frac{\Sigma X^2 - [(\Sigma X)^2/n]}{n-1}$$

Formula for the variance for grouped data:

$$s^2 = \frac{\Sigma f \cdot X_m^2 - [(\Sigma f \cdot X_m)^2/n]}{n-1}$$

Formula for the standard deviation for population data:

$$\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$$

Formula for the standard deviation for sample data (shortcut formula):

$$s = \sqrt{\frac{\Sigma X^2 - [(\Sigma X)^2/n]}{n-1}}$$

Formula for the standard deviation for grouped data:

$$s = \sqrt{\frac{\Sigma f \cdot X_m^2 - [(\Sigma f \cdot X_m)^2/n]}{n-1}}$$

Formula for the coefficient of variation:

$$\text{CVar} = \frac{s}{\overline{X}} \cdot 100\% \qquad \text{or} \qquad \text{CVar} = \frac{\sigma}{\mu} \cdot 100\%$$

Range rule of thumb:

$$s \approx \frac{\text{range}}{4}$$

Expression for Chebyshev's theorem: The proportion of values from a data set that will fall within *k* standard deviations of the mean will be at least

$$1 - \frac{1}{k^2}$$

where *k* is a number greater than 1.

Formula for the *z* score (standard score):

$$z = \frac{X-\mu}{\sigma} \qquad \text{or} \qquad z = \frac{X-\overline{X}}{s}$$

Formula for the cumulative percentage:

$$\textbf{Cumulative \%} = \frac{\textbf{cumulative frequency}}{n} \cdot 100\%$$

Formula for the percentile rank of a value *X*:

$$\textbf{Percentile} = \frac{\left(\begin{array}{c}\textbf{number of values}\\ \textbf{below } X + 0.5\end{array}\right)}{\begin{array}{c}\textbf{total number}\\ \textbf{of values}\end{array}} \cdot 100\%$$

Formula for finding a value corresponding to a given percentile:

$$c = \frac{n \cdot p}{100}$$

Formula for interquartile range:

$$\textbf{IQR} = Q_3 - Q_1$$

Review Exercises

1. The following data represent the number of listeners (in thousands) of 15 radio stations in the 6:00 to 9:00 A.M. time slot in Pittsburgh.

229, 182, 129, 112, 122, 93, 97, 114, 95, 114, 60, 89, 75, 70, 68

Source: Arbitron Inc.

Find each of these.

a. Mean
b. Median
c. Mode
d. Midrange
e. Range
f. Variance
g. Standard deviation

2. These data represent the area in square miles of major islands in the Caribbean Sea and the Mediterranean Sea.

Caribbean Sea			Mediterranean Sea	
108	926	436	1,927	1,411
75	100	3,339	229	95
5,382	171	116	3,189	540
2,300	290	1,864	3,572	9,301
166	687	59	86	9,926
42,804	4,244	134		
29,389				

Source: The World Almanac and Book of Facts.

Find each of these.

a. Mean
b. Median
c. Mode
d. Midrange
e. Range
f. Variance
g. Standard deviation

Are the averages and variations of the areas approximately equal?

3. Twelve batteries were tested to see how many hours they would last. The frequency distribution is shown here.

Hours	Frequency
1–3	1
4–6	4
7–9	5
10–12	1
13–15	1

Find each of these.

a. Mean
b. Modal class
c. Variance
d. Standard deviation

4. The following data represent the number of seconds it took 20 students to find information from the Internet on a personal computer.

Class	Frequency
34–38	4
39–43	6
44–48	3
49–53	4
54–58	3

Find each of these.

a. Mean
b. Modal class
c. Variance
d. Standard deviation

5. Shown here is a frequency distribution for the rise in tides at 30 selected locations in the United States.

Rise in tides (inches)	Frequency
12.5–27.5	6
27.5–42.5	3
42.5–57.5	5
57.5–72.5	8
72.5–87.5	6
87.5–102.5	2

Find each of these.

a. Mean
b. Modal class
c. Variance
d. Standard deviation

6. The fuel capacity in gallons of 50 randomly selected 1995 cars is shown here.

Class	Frequency
10–12	6
13–15	4
16–18	14
19–21	15
22–24	8
25–27	2
28–30	1
	50

Find each of these.

a. Mean
b. Modal class
c. Variance
d. Standard deviation

7. In a dental survey of third-grade students, this distribution was obtained for the number of cavities found. Find the average number of cavities for the class. Use the weighted mean.

Number of students	Number of cavities
12	0
8	1
5	2
5	3

8. An investor calculated these percentages of each of three stock investments with payoffs as shown. Find the average payoff. Use the weighted mean.

Stock	Percent	Payoff
A	30	$10,000
B	50	3,000
C	20	1,000

9. In an advertisement, a transmission service center stated that the average years of service of its employees were 13. The distribution is shown here. Using the weighted mean, calculate the correct average.

Number of employees	Years of service
8	3
1	6
1	30

10. If the average number of textbooks in professors' offices is 16, the standard deviation is 5, and the average age of the professors is 43, with a standard deviation of 8, which data set is more variable?

11. A survey of bookstores showed that the average number of magazines carried is 56, with a standard deviation of 12. The same survey showed that the average length of time each store had been in business was 6 years, with a standard deviation of 2.5 years. Which is more variable, the number of magazines or the number of years?

12. The number of previous jobs held by each of six applicants is shown here.

2, 4, 5, 6, 8, 9

a. Find the percentile for each value.
b. What value corresponds to the 30th percentile?
c. Construct a boxplot and comment on the nature of the distribution.

13. The salaries (in millions of dollars) for 29 NFL teams for the 1999–2000 season are given in this frequency distribution.

Class limits	Frequency
39.9–42.8	2
42.9–45.8	2
45.9–48.8	5
48.9–51.8	5
51.9–54.8	12
54.9–57.8	3

Source: www.NFL.com

a. Construct a percentile graph.
b. Find the values that correspond to the 35th, 65th, and 85th percentiles.
c. Find the percentile of values 44, 48, and 54.

14. Check each data set for outliers.

a. 506, 511, 517, 514, 400, 521
b. 3, 7, 9, 6, 8, 10, 14, 16, 20, 12
c. 14, 18, 27, 26, 19, 13, 5, 25
d. 112, 157, 192, 116, 153, 129, 131

15. A survey of car rental agencies shows that the average cost of a car rental is $0.32 per mile. The standard deviation is $0.03. Using Chebyshev's theorem, find the range in which at least 75% of the data values will fall.

16. The average cost of a certain type of seed per acre is $42. The standard deviation is $3. Using Chebyshev's theorem, find the range in which at least 88.89% of the data values will fall.

17. The average labor charge for automobile mechanics is $54 per hour. The standard deviation is $4. Find the minimum percentage of data values that will fall within the range of $48 to $60. Use Chebyshev's theorem.

18. For a certain type of job, it costs a company an average of $231 to train an employee to perform the task. The standard deviation is $5. Find the minimum percentage of data values that will fall in the range of $219 to $243. Use Chebyshev's theorem.

19. The average delivery charge for a refrigerator is $32. The standard deviation is $4. Find the minimum percentage of data values that will fall in the range of $20 to $44. Use Chebyshev's theorem.

20. Which of these exam grades has a better relative position?

a. A grade of 82 on a test with $\overline{X} = 85$ and $s = 6$.
b. A grade of 56 on a test with $\overline{X} = 60$ and $s = 5$.

21. The data shown here represent the number of hours that 12 part-time employees at a toy store worked during the weeks before and after Christmas. Construct two boxplots and compare the distributions.

Before	38	16	18	24	12	30	35	32	31	30	24	35
After	26	15	12	18	24	32	14	18	16	18	22	12

22. The mean of the times it takes a commuter to get to work in Baltimore is 29.7 minutes. If the standard deviation is 6 minutes, within what limits would you expect approximately 68% of the times to fall? Assume the distribution is approximately bell-shaped.

Statistics Today

How Long Are You Delayed by Road Congestion?—Revisited

The average number of hours per year that a driver is delayed by road congestion is listed here.

Los Angeles	56
Atlanta	53
Seattle	53
Houston	50
Dallas	46
Washington	46
Austin	45
Denver	45
St. Louis	44
Orlando	42
U.S. average	36

Source: Texas Transportation Institute.

By making comparisons using averages, you can see that drivers in these 10 cities are delayed by road congestion more than the national average.

Data Analysis

A Data Bank is found in Appendix D, or on the World Wide Web by following links from www.mhhe.com/math/stat/bluman/

1. From the Data Bank, choose one of the following variables: age, weight, cholesterol level, systolic pressure, IQ, or sodium level. Select at least 30 values, and find the mean, median, mode, and midrange. State which measurement of central tendency best describes the average and why.
2. Find the range, variance, and standard deviation for the data selected in Exercise 1.
3. From the Data Bank, choose 10 values from any variable, construct a boxplot, and interpret the results.
4. Randomly select 10 values from the number of suspensions in the local school districts in southwestern Pennsylvania in Data Set V in Appendix D. Find the mean, median, mode, range, variance, and standard deviation of the number of suspensions by using the Pearson coefficient of skewness.
5. Using the data from Data Set VII in Appendix D, find the mean, median, mode, range, variance, and standard deviation of the acreage owned by the municipalities. Comment on the skewness of the data, using the Pearson coefficient of skewness.

Chapter Quiz

Determine whether each statement is true or false. If the statement is false, explain why.

1. When the mean is computed for individual data, all values in the data set are used.
2. The mean cannot be found for grouped data when there is an open class.
3. A single, extremely large value can affect the median more than the mean.
4. One-half of all the data values will fall above the mode, and one-half will fall below the mode.
5. In a data set, the mode will always be unique.
6. The range and midrange are both measures of variation.
7. One disadvantage of the median is that it is not unique.
8. The mode and midrange are both measures of variation.
9. If a person's score on an exam corresponds to the 75th percentile, then that person obtained 75 correct answers out of 100 questions.

Select the best answer.

10. What is the value of the mode when all values in the data set are different?

    a. 0
    b. 1
    c. There is no mode.
    d. It cannot be determined unless the data values are given.

11. When data are categorized as, for example, places of residence (rural, suburban, urban), the most appropriate measure of central tendency is the

    a. Mean *c.* Mode
    b. Median *d.* Midrange

12. P_{50} corresponds to

    a. Q_2
    b. D_5
    c. IQR
    d. Midrange

13. Which is not part of the five-number summary?

    a. Q_1 and Q_3
    b. The mean
    c. The median
    d. The smallest and the largest data values

14. A statistic that tells the number of standard deviations a data value is above or below the mean is called

    a. A quartile
    b. A percentile
    c. A coefficient of variation
    d. A z score

15. When a distribution is bell-shaped, approximately what percentage of data values will fall within 1 standard deviation of the mean?

    a. 50%
    b. 68%
    c. 95%
    d. 99.7%

Complete these statements with the best answer.

16. A measure obtained from sample data is called a(n) __________.
17. Generally, Greek letters are used to represent __________, and Roman letters are used to represent __________.
18. The positive square root of the variance is called the __________.
19. The symbol for the population standard deviation is __________.
20. When the sum of the lowest data value and the highest data value is divided by 2, the measure is called __________.
21. If the mode is to the left of the median and the mean is to the right of the median, then the distribution is __________ skewed.
22. An extremely high or extremely low data value is called a(n) __________.
23. The number of highway miles per gallon of the 10 worst vehicles is shown.

    12 15 13 14 15 16 17 16 17 18

    Source: Pittsburgh Post Gazette.

    Find each of these.

    a. Mean
    b. Median
    c. Mode
    d. Midrange
    e. Range
    f. Variance
    g. Standard deviation

24. The distribution of the number of errors that 10 students made on a typing test is shown.

Errors	Frequency
0–2	1
3–5	3
6–8	4
9–11	1
12–14	1

Find each of these.

a. Mean
b. Modal class
c. Variance
d. Standard deviation

25. Shown here is a frequency distribution for the number of inches of rain received in 1 year in 25 selected cities in the United States.

Number of inches	Frequency
5.5–20.5	2
20.5–35.5	3
35.5–50.5	8
50.5–65.5	6
65.5–80.5	3
80.5–95.5	3

Find each of these.

a. Mean
b. Modal class
c. Variance
d. Standard deviation

26. A survey of 36 selected recording companies showed these numbers of days that it took to receive a shipment from the day it was ordered.

Days	Frequency
1–3	6
4–6	8
7–9	10
10–12	7
13–15	0
16–18	5

Find each of these.

a. Mean
b. Modal class
c. Variance
d. Standard deviation

27. In a survey of third-grade students, this distribution was obtained for the number of "best friends" each had.

Number of students	Number of best friends
8	1
6	2
5	3
3	0

Find the average number of best friends for the class. Use the weighted mean.

28. In an advertisement, a retail store stated that its employees averaged 9 years of service. The distribution is shown here.

Number of employees	Years of service
8	2
2	6
3	10

Using the weighted mean, calculate the correct average.

29. The average number of newspapers for sale in an airport newsstand is 12, and the standard deviation is 4. The average age of the pilots is 37 years, with a standard deviation of 6 years. Which data set is more variable?

30. A survey of grocery stores showed that the average number of brands of toothpaste carried was 16, with a standard deviation of 5. The same survey showed the average length of time each store was in business was 7 years, with a standard deviation of 1.6 years. Which is more variable, the number of brands or the number of years?

31. A student scored 76 on a general science test where the class mean and standard deviation were 82 and 8, respectively; he also scored 53 on a psychology test where the class mean and standard deviation were 58 and 3, respectively. In which class was his relative position higher?

32. Which score has the highest relative position?

a. $X = 12$ $\overline{X} = 10$ $s = 4$
b. $X = 170$ $\overline{X} = 120$ $s = 32$
c. $X = 180$ $\overline{X} = 60$ $s = 8$

33. The number of square feet (in millions) of 8 of the largest malls in southwestern Pennsylvania is shown.

1	0.9	1.3	0.8
1.4	0.77	0.7	1.2

Source: International Council of Shopping Centers.

a. Find the percentile for each value.
b. What value corresponds to the 40th percentile?
c. Construct a boxplot and comment on the nature of the distribution.

34. On a philosophy comprehensive exam, this distribution was obtained from 25 students.

Score	Frequency
40.5–45.5	3
45.5–50.5	8
50.5–55.5	10
55.5–60.5	3
60.5–65.5	1

a. Construct a percentile graph.
b. Find the values that correspond to the 22nd, 78th, and 99th percentiles.
c. Find the percentiles of the values 52, 43, and 64.

35. The first column of these data represents the prebuy gas price of a rental car, and the second column represents the price charged if the car is returned without refilling the gas tank for a selected car rental company. Draw two boxplots for the data and compare the distributions.

Prebuy cost	No prebuy cost
$1.55	$3.80
1.54	3.99
1.62	3.99
1.65	3.85
1.72	3.99
1.63	3.95
1.65	3.94
1.72	4.19
1.45	3.84
1.52	3.94

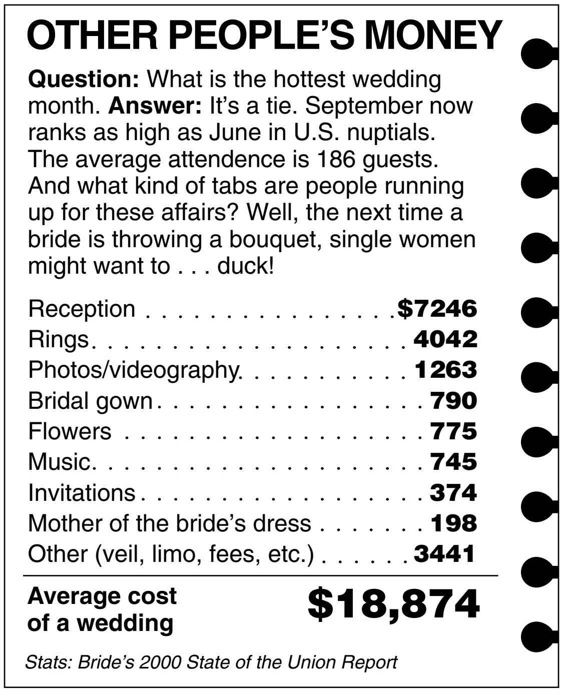
Source: USA TODAY.

36. The average national SAT score is 1019. If we assume a bell-shaped distribution and a standard deviation equal to 110, what percentage of scores would you expect to fall above 1129? Above 799?

Source: N.Y. Times Almanac, 2002.

Critical Thinking Challenges

1. Averages give us information to help us to see where we stand and enable us to make comparisons. Here is a study on the average cost of a wedding. What type of average—mean, median, mode, or midrange—might have been used for each category?



OTHER PEOPLE'S MONEY

Question: What is the hottest wedding month. **Answer:** It's a tie. September now ranks as high as June in U.S. nuptials. The average attendence is 186 guests. And what kind of tabs are people running up for these affairs? Well, the next time a bride is throwing a bouquet, single women might want to . . . duck!

Reception	**$7246**
Rings	**4042**
Photos/videography	**1263**
Bridal gown	**790**
Flowers	**775**
Music	**745**
Invitations	**374**
Mother of the bride's dress	**198**
Other (veil, limo, fees, etc.)	**3441**
Average cost of a wedding	**$18,874**

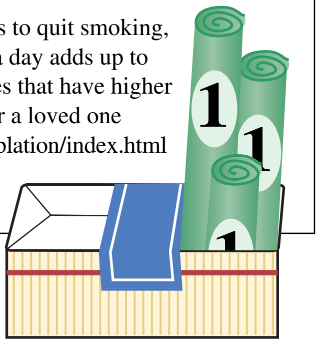
Stats: Bride's 2000 State of the Union Report

Source: Reprinted with permission from the September 2001 Reader's Digest. Copyright © 2001 by The Reader's Digest Assn., Inc.

2. This article states that the average yearly cost of smoking a pack of cigarettes a day is $1190. Find the average cost of a pack of cigarettes in your area, and compute the cost per day for 1 year. Compare your answer with the one in the article.

Burning Through the Cash

Everyone knows the health-related reasons to quit smoking, so here's an economic argument: A pack a day adds up to $1190 a year on average; it's more in states that have higher taxes on tobacco. To calculate what you or a loved one spends, visit ashline.org/ASH/quit/contemplation/index.html and try out their smoker's calculator.
You'll be stunned.



Source: Reprinted with permission from the April 2002 Reader's Digest. Copyright © 2002 by The Reader's Digest Assn., Inc.

3. The table shows the median ages of residents for the 10 oldest states and the 10 youngest states of the United States including Washington, D.C. Explain why the median is used instead of the mean.

10 Oldest			10 Youngest		
Rank	**State**	**Median age**	**Rank**	**State**	**Median age**
1	West Virginia	38.9	51	Utah	27.1
2	Florida	38.7	50	Texas	32.3
3	Maine	38.6	49	Alaska	32.4
4	Pennsylvania	38.0	48	Idaho	33.2
5	Vermont	37.7	47	California	33.3
6	Montana	37.5	46	Georgia	33.4
7	Connecticut	37.4	45	Mississippi	33.8
8	New Hampshire	37.1	44	Louisiana	34.0
9	New Jersey	36.7	43	Arizona	34.2
10	Rhode Island	36.7	42	Colorado	34.3

Source: U.S. Census Bureau.

Data Projects

Where appropriate, use MINITAB, the TI-83 Plus, the TI-84 Plus, or a computer program of your choice to complete the following exercises.

1. Select a variable and collect about 10 values for two groups. (For example, you may want to ask 10 men how many cups of coffee they drink per day and 10 women the same question.)

a. Define the variable.
b. Define the populations.
c. Describe how the samples were selected.
d. Write a paragraph describing the similarities and differences between the two groups, using appropriate descriptive statistics such as means, standard deviations, and so on.

2. Collect data consisting of at least 30 values.

a. State the purpose of the project.
b. Define the population.
c. State how the sample was selected.
d. Using appropriate descriptive statistics, write a paragraph summarizing the data.

You may use the following websites to obtain raw data:

Visit the data sets at the book's website found at http://www.mhhe.com/math/stat/bluman. Click on the 6th edition.

http://lib.stat.cmu.edu/DASL

http://www.statcan.ca

Answers to Applying the Concepts

Section 3–2 Teacher Salaries

1. The sample mean is $22,921.67, the sample median is $16,500, and the sample mode is $11,000. If you work for the school board and do not want to raise salaries, you could say that the average teacher salary is $22,921.67.
2. If you work for the teachers' union and want a raise for the teachers, either the sample median of $16,500 or the sample mode of $11,000 would be a good measure of center to report.
3. The outlier is $107,000. With the outlier removed, the sample mean is $15,278.18, the sample median is $16,400, and the sample mode is still $11,000. The mean is greatly affected by the outlier and allows the school board to report an average teacher salary that is not representative of a "typical" teacher salary.
4. If the salaries represented every teacher in the school district, the averages would be parameters, since we have data from the entire population.
5. The mean can be misleading in the presence of outliers, since it is greatly affected by these extreme values.
6. Since the mean is greater than both the median and the mode, the distribution is skewed to the right (positively skewed).

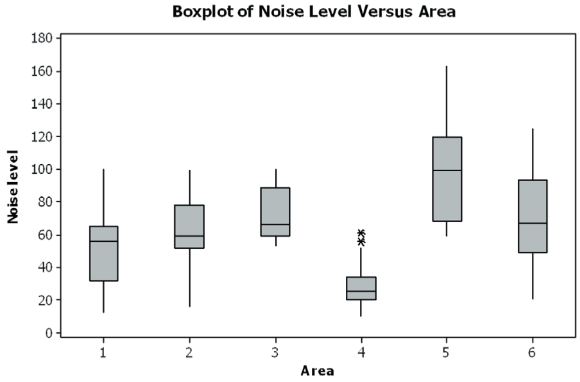
Section 3–3 Blood Pressure

1. Chebyshev's theorem does not work for scores within 1 standard deviation of the mean.
2. At least 75% (900) of the normotensive men will fall in the interval 105–141 mm Hg.
3. About 95% (1330) of the normotensive women have diastolic blood pressures between 62 and 90 mm Hg. About 95% (1235) of the hypertensive women have diastolic blood pressures between 68 and 108 mm Hg.
4. About 95% (1140) of the normotensive men have systolic blood pressures between 105 and 141 mm Hg. About 95% (1045) of the hypertensive men have systolic blood pressures between 119 and 187 mm Hg. These two ranges do overlap.

Section 3–4 Determining Dosages

1. The quartiles could be used to describe the data results.
2. Since there are 10 mice in the upper quartile, this would mean that 4 of them survived.
3. The percentiles would give us the position of a single mouse with respect to all other mice.
4. The quartiles divide the data into four groups of equal size.
5. Standard scores would give us the position of a single mouse with respect to the mean time until the onset of sepsis.

Section 3–5 The Noisy Workplace



From this boxplot, we see that about 25% of the readings in area 5 are above the safe hearing level of 120 decibels. Those workers in area 5 should definitely have protective ear wear. One of the readings in area 6 is above the safe hearing level. It might be a good idea to provide protective ear wear to those workers in area 6 as well. Areas 1–4 appear to be "safe" with respect to hearing level, with area 4 being the safest.