

3

Describing Data:

Numerical Measures



U. S. housing prices reached their peak in July of 2006. Based on the data given in the table, showing the ratio of the prices reported in a recent month to the July 2006 figure, find the mean, median, and mode of these statistics. (Exercise 83, Goal 1)

GOALS

When you have completed this chapter you will be able to:

- 1 Calculate the *arithmetic mean*, *weighted mean*, *median*, *mode*, and *geometric mean*.
- 2 Explain the characteristics, uses, advantages, and disadvantages of each *measure of location*.
- 3 Identify the position of the mean, median, and mode for both *symmetric* and *skewed distributions*.
- 4 Compute and interpret the *range*, *mean deviation*, *variance*, and *standard deviation*.
- 5 Understand the characteristics, uses, advantages, and disadvantages of each *measure of dispersion*.
- 6 Understand *Chebyshev's theorem* and the *Empirical Rule* as they relate to a set of observations.



Statistics in Action

Did you ever meet the “average” American? Well, his name is Robert (that is the nominal level of measurement), he is 31 years old (that is the ratio level), he is 69.5 inches tall (again the ratio level of measurement), weighs 172 pounds, wears a size 9½ shoe, has a 34-inch waist, and wears a size 40 suit. In addition, the average man eats 4 pounds of potato chips, watches 2,567 hours of TV, and eats 26 pounds of bananas each year and also sleeps 7.7 hours per night.

The average American woman is 5' 4" tall and weighs 140 pounds, while the average American model is 5' 11" tall and weighs 117 pounds. On any given day, almost half of the women in the United States are on a diet. Idolized in the 1950s, Marilyn Monroe would be considered overweight by today's standards. She fluctuated between a size 14 and 18 dress, and was a healthy and attractive woman.

Introduction

Chapter 2 began our study of descriptive statistics. To transform a mass of raw data into a meaningful form, we organized quantitative data into a frequency distribution and portrayed it graphically in a histogram. We also looked at other graphical techniques such as pie charts to portray qualitative data and frequency polygons to portray quantitative data.



This chapter is concerned with two numerical ways of describing quantitative data, namely, **measures of location** and **measures of dispersion**. Measures of location are often referred to as **averages**. The purpose of a measure of location is to pinpoint the center of a set of values.

You are familiar with the concept of an average. An average is a measure of location that shows the central value of the data. Averages appear daily on TV, in the newspaper, and other journals. Here are some examples:

- The average U.S. home changes ownership every 11.8 years.
- An American receives an average of 568 pieces of mail per year.
- The average American home has more TV sets than people. There are 2.73 TV sets and 2.55 people in the typical home.
- The average hourly earnings for civilian workers in the United States is \$19.29 per hour.
- The average American couple spends \$28,732 for their wedding, while their budget is 50 percent less. This does not include the cost of a honeymoon or engagement ring.
- In Chicago the mean high temperature is 84 degrees in July and 31 degrees in January. The mean amount of precipitation is 3.80 inches in July and 1.90 inches in January.

If we consider only the measures of location in a set of data, or if we compare several sets of data using central values, we may draw an erroneous conclusion. In addition to the measures of location, we should consider the **dispersion**—often called the *variation* or the *spread*—in the data. As an illustration, suppose the average annual income of executives for Internet-related companies is \$80,000, and the average income for executives in pharmaceutical firms is also \$80,000. If we looked only at the average incomes, we might wrongly conclude that the two salary distributions are identical or nearly identical. A look at the salary ranges indicates that this conclusion is not correct. The salaries for the executives in the Internet firms range from \$70,000 to \$90,000, but salaries for the marketing executives in pharmaceuticals range from \$40,000 to \$120,000. Thus, we conclude that although the average salaries are the same for the two industries, there is much more spread or dispersion in salaries for the pharmaceutical executives. To describe the dispersion we will consider the range, the mean deviation, the variance, and the standard deviation.

We begin by discussing measures of location. There is not just one measure of location; in fact, there are many. We will consider five: the arithmetic mean, the weighted mean, the median, the mode, and the geometric mean. The arithmetic mean is the most widely used and widely reported measure of location. We study the mean as both a population parameter and a sample statistic.

The Population Mean

Many studies involve all the values in a population. For example, there are 39 exits on I-75 through the state of Kentucky. The mean distance between these state exits is 4.76 miles. This is an example of a population parameter because we have studied the distance between *all* the exits. There are 12 sales associates employed at the Reynolds Road outlet of Carpets by Otto. The mean amount of commission they earned last month was \$1,345. This is a population value because we considered the commission of *all* the sales associates. Other examples of a population mean would be: the mean closing price for Johnson & Johnson stock for the last 5 days is \$61.75; the mean annual rate of return for the last 10 years for Berger Funds is 8.67 percent; and the mean number of hours of overtime worked last week by the six welders in the welding department of Butts Welding, Inc., is 6.45 hours.

For raw data, that is, data that has not been grouped in a frequency distribution, the population mean is the sum of all the values in the population divided by the number of values in the population. To find the population mean, we use the following formula.

$$\text{Population mean} = \frac{\text{Sum of all the values in the population}}{\text{Number of values in the population}}$$

Instead of writing out in words the full directions for computing the population mean (or any other measure), it is more convenient to use the shorthand symbols of mathematics. The mean of a population using mathematical symbols is:

POPULATION MEAN

$$\mu = \frac{\sum X}{N}$$

[3-1]

where:

μ represents the population mean. It is the Greek lowercase letter “mu.”

N is the number of values in the population.

X represents any particular value.

\sum is the Greek capital letter “sigma” and indicates the operation of adding.

$\sum X$ is the sum of the X values in the population.

Any measurable characteristic of a population is called a **parameter**. The mean of a population is a parameter.

PARAMETER A characteristic of a population.

Example

There are 12 automobile manufacturing companies in the United States. Listed below is the number of patents granted by the United States government to each company in a recent year.

Company	Number of Patents Granted	Company	Number of Patents Granted
General Motors	511	Mazda	210
Nissan	385	Chrysler	97
Daimler	275	Porsche	50
Toyota	257	Mitsubishi	36
Honda	249	Volvo	23
Ford	234	BMW	13

Is this information a sample or a population? What is the arithmetic mean number of patents granted?

Solution

This is a population because we are considering all the automobile manufacturing companies obtaining patents. We add the number of patents for each of the 12 companies. The total number of patents for the 12 companies is 2,340. To find the arithmetic mean, we divide this total by 12. So the arithmetic mean is 195, found by $2,340/12$. From formula (3-1):

$$\mu = \frac{511 + 385 + \cdots + 13}{12} = \frac{2340}{12} = 195$$

How do we interpret the value of 195? The typical number of patents received by an automobile manufacturing company is 195. Because we considered all the companies receiving patents, this value is a population parameter.

The Sample Mean

As explained in Chapter 1, we often select a sample from the population to find something about a specific characteristic of the population. The quality assurance department, for example, needs to be assured that the ball bearings being produced have an acceptable outside diameter. It would be very expensive and time consuming to check the outside diameter of all the bearings produced. Therefore, a sample of five bearings is selected and the mean outside diameter of the five bearings is calculated to estimate the mean diameter of all the bearings.

For raw data, that is, ungrouped data, *the mean is the sum of all the sampled values divided by the total number of sampled values*. To find the mean for a sample:

$$\text{Sample mean} = \frac{\text{Sum of all the values in the sample}}{\text{Number of values in the sample}}$$

The mean of a sample and the mean of a population are computed in the same way, but the shorthand notation used is different. The formula for the mean of a *sample* is:

SAMPLE MEAN

$$\bar{X} = \frac{\sum X}{n}$$

[3-2]

where:

\bar{X} is the sample mean. It is read “X bar.”

n is the number of values in the sample.

The mean of a sample, or any other measure based on sample data, is called a **statistic**. If the mean outside diameter of a sample of five ball bearings is 0.625 inches, this is an example of a statistic.

STATISTIC A characteristic of a sample.

Example

SunCom is studying the number of minutes used by clients in a particular cell phone rate plan. A random sample of 12 clients showed the following number of minutes used last month.

90	77	94	89	119	112
91	110	92	100	113	83

What is the arithmetic mean number of minutes used?

Mean of ungrouped
sample data

Solution

Using formula (3-2), the sample mean is:

$$\text{Sample mean} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$$

$$\bar{X} = \frac{\sum X}{n} = \frac{90 + 77 + \cdots + 83}{12} = \frac{1170}{12} = 97.5$$

The arithmetic mean number of minutes used last month by the sample of cell phone users is 97.5 minutes.

Properties of the Arithmetic Mean

The arithmetic mean is a widely used measure of location. It has several important properties:

1. **Every set of interval- or ratio-level data has a mean.** Recall from Chapter 1 that ratio-level data include such data as ages, incomes, and weights, with the distance between numbers being constant.
2. **All the values are included in computing the mean.**
3. **The mean is unique.** That is, there is only one mean in a set of data. Later in the chapter we will discover an average that might appear twice, or more than twice, in a set of data.
4. **The sum of the deviations of each value from the mean is zero.** Expressed symbolically:

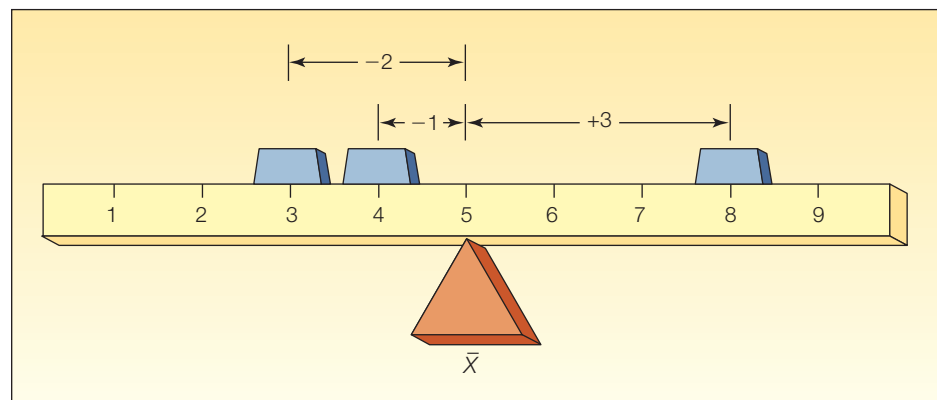
$$\sum(X - \bar{X}) = 0$$

As an example, the mean of 3, 8, and 4 is 5. Then:

$$\begin{aligned}\sum(X - \bar{X}) &= (3 - 5) + (8 - 5) + (4 - 5) \\ &= -2 + 3 - 1 \\ &= 0\end{aligned}$$

Mean as a balance point

Thus, we can consider the mean as a balance point for a set of data. To illustrate, we have a long board with the numbers 1, 2, 3, . . . , 9 evenly spaced on it. Suppose three bars of equal weight were placed on the board at numbers 3, 4, and 8, and the balance point was set at 5, the mean of the three numbers. We would find that the board is balanced perfectly! The deviations below the mean (-3) are equal to the deviations above the mean ($+3$). Shown schematically:



Mean unduly affected by unusually large or small values

The mean does have a weakness. Recall that the mean uses the value of every item in a sample, or population, in its computation. If one or two of these values are

either extremely large or extremely small compared to the majority of data, the mean might not be an appropriate average to represent the data. For example, suppose the annual incomes of a small group of stockbrokers at Merrill Lynch are \$62,900, \$61,600, \$62,500, \$60,800, and \$1,200,000. The mean income is \$289,560. Obviously, it is not representative of this group, because all but one broker has an income in the \$60,000 to \$63,000 range. One income (\$1.2 million) is unduly affecting the mean.

Self-Review 3–1



- The annual incomes of a sample of middle-management employees at Westinghouse are: \$62,900, \$69,100, \$58,300, and \$76,800.
 - Give the formula for the sample mean.
 - Find the sample mean.
 - Is the mean you computed in (b) a statistic or a parameter? Why?
 - What is your best estimate of the population mean?
- All the students in advanced Computer Science 411 are a population. Their course grades are 92, 96, 61, 86, 79, and 84.
 - Give the formula for the population mean.
 - Compute the mean course grade.
 - Is the mean you computed in (b) a statistic or a parameter? Why?

Exercises

connect™

The answers to the odd-numbered exercises are at the end of the book.

- Compute the mean of the following population values: 6, 3, 5, 7, 6.
- Compute the mean of the following population values: 7, 5, 7, 3, 7, 4.
- Compute the mean of the following sample values: 5, 9, 4, 10.
 - Show that $\Sigma(X - \bar{X}) = 0$.
- Compute the mean of the following sample values: 1.3, 7.0, 3.6, 4.1, 5.0.
 - Show that $\Sigma(X - \bar{X}) = 0$.
- Compute the mean of the following sample values: 16.25, 12.91, 14.58.
- Compute the mean hourly wage paid to carpenters who earned the following hourly wages: \$15.40, \$20.10, \$18.75, \$22.76, \$30.67, \$18.00.

For Exercises 7–10, (a) compute the arithmetic mean and (b) indicate whether it is a statistic or a parameter.

- There are 10 salespeople employed by Midtown Ford. The number of new cars sold last month by the respective salespeople were: 15, 23, 4, 19, 18, 10, 10, 8, 28, 19.
- The accounting department at a mail-order company counted the following numbers of incoming calls per day to the company's toll-free number during the first 7 days in May: 14, 24, 19, 31, 36, 26, 17.
- The Cambridge Power and Light Company selected a random sample of 20 residential customers. Following are the amounts, to the nearest dollar, the customers were charged for electrical service last month:

54	48	58	50	25	47	75	46	60	70
67	68	39	35	56	66	33	62	65	67

- The Human Relations Director at Ford began a study of the overtime hours in the Inspection Department. A sample of 15 workers showed they worked the following number of overtime hours last month.

13	13	12	15	7	15	5	12
6	7	12	10	9	13	12	

- AAA Heating and Air Conditioning completed 30 jobs last month with a mean revenue of \$5,430 per job. The president wants to know the total revenue for the month. Based on the limited information, can you compute the total revenue? What is it?

12. A large pharmaceutical company hires business administration graduates to sell its products. The company is growing rapidly and dedicates only one day of sales training for new salespeople. The company's goal for new salespeople is \$10,000 per month. The goal is based on the current mean sales for the entire company, which is \$10,000 per month. After reviewing the retention rates of new employees, the company finds that only 1 in 10 new employees stays longer than three months. Comment on using the current mean sales per month as a sales goal for new employees. Why do new employees leave the company?

The Weighted Mean

The weighted mean is a special case of the arithmetic mean. It occurs when there are several observations of the same value. To explain, suppose the nearby Wendy's Restaurant sold medium, large, and Biggie-sized soft drinks for \$.90, \$1.25, and \$1.50, respectively. Of the last 10 drinks sold, 3 were medium, 4 were large, and 3 were Biggie-sized. To find the mean price of the last 10 drinks sold, we could use formula (3-2).

$$\bar{X} = \frac{\$.90 + \$.90 + \$.90 + \$1.25 + \$1.25 + \$1.25 + \$1.25 + \$1.50 + \$1.50 + \$1.50}{10}$$

$$\bar{X} = \frac{\$12.20}{10} = \$1.22$$

The mean selling price of the last 10 drinks is \$1.22.

An easier way to find the mean selling price is to determine the weighted mean. That is, we multiply each observation by the number of times it happens. We will refer to the weighted mean as \bar{X}_w . This is read "X bar sub w."

$$\bar{X}_w = \frac{3(\$0.90) + 4(\$1.25) + 3(\$1.50)}{10} = \frac{\$12.20}{10} = \$1.22$$

In this case the weights are frequency counts. However, any measure of importance could be used as a weight. In general the weighted mean of a set of numbers designated $X_1, X_2, X_3, \dots, X_n$ with the corresponding weights $w_1, w_2, w_3, \dots, w_n$ is computed by:

WEIGHTED MEAN

$$\bar{X}_w = \frac{w_1X_1 + w_2X_2 + w_3X_3 + \dots + w_nX_n}{w_1 + w_2 + w_3 + \dots + w_n}$$

[3-3]

This may be shortened to:

$$\bar{X}_w = \frac{\Sigma(wX)}{\Sigma w}$$

Note that the denominator of a weighted mean is always the sum of the weights.

Example

The Carter Construction Company pays its hourly employees \$16.50, \$19.00, or \$25.00 per hour. There are 26 hourly employees, 14 of which are paid at the \$16.50 rate, 10 at the \$19.00 rate, and 2 at the \$25.00 rate. What is the mean hourly rate paid the 26 employees?

Solution

To find the mean hourly rate, we multiply each of the hourly rates by the number of employees earning that rate. From formula (3-3), the mean hourly rate is

$$\bar{X}_w = \frac{14(\$16.50) + 10(\$19.00) + 2(\$25.00)}{14 + 10 + 2} = \frac{\$471.00}{26} = \$18.1154$$

The weighted mean hourly wage is rounded to \$18.12.

Self-Review 3–2

Springers sold 95 Antonelli men's suits for the regular price of \$400. For the spring sale the suits were reduced to \$200 and 126 were sold. At the final clearance, the price was reduced to \$100 and the remaining 79 suits were sold.

- What was the weighted mean price of an Antonelli suit?
- Springers paid \$200 a suit for the 300 suits. Comment on the store's profit per suit if a salesperson receives a \$25 commission for each one sold.

Exercises

connect™

- In June an investor purchased 300 shares of Oracle (an information technology company) stock at \$20 per share. In August she purchased an additional 400 shares at \$25 per share. In November she purchased an additional 400 shares, but the stock declined to \$23 per share. What is the weighted mean price per share?
- The Bookstall, Inc., is a specialty bookstore concentrating on used books sold via the Internet. Paperbacks are \$1.00 each, and hardcover books are \$3.50. Of the 50 books sold last Tuesday morning, 40 were paperback and the rest were hardcover. What was the weighted mean price of a book?
- The Loris Healthcare System employs 200 persons on the nursing staff. Fifty are nurse's aides, 50 are practical nurses, and 100 are registered nurses. Nurse's aides receive \$8 an hour, practical nurses \$15 an hour, and registered nurses \$24 an hour. What is the weighted mean hourly wage?
- Andrews and Associates specialize in corporate law. They charge \$100 an hour for researching a case, \$75 an hour for consultations, and \$200 an hour for writing a brief. Last week one of the associates spent 10 hours consulting with her client, 10 hours researching the case, and 20 hours writing the brief. What was the weighted mean hourly charge for her legal services?

The Median

We have stressed that, for data containing one or two very large or very small values, the arithmetic mean may not be representative. The center for such data can be better described by a measure of location called the **median**.

To illustrate the need for a measure of location other than the arithmetic mean, suppose you are seeking to buy a condominium in Palm Aire. Your real estate agent says that the typical price of the units currently available is \$110,000. Would you still want to look? If you had budgeted your maximum purchase price at \$75,000, you might think they are out of your price range. However, checking the prices of the individual units might change your mind. They are \$60,000, \$65,000, \$70,000, and \$80,000, and a superdeluxe penthouse costs \$275,000. The arithmetic mean price is \$110,000, as the real estate agent reported, but one price (\$275,000) is pulling the arithmetic mean upward, causing it to be an unrepresentative average. It does seem that a price around \$70,000 is a more typical or representative average, and it is. In cases such as this, the median provides a more valid measure of location.

MEDIAN The midpoint of the values after they have been ordered from the smallest to the largest, or the largest to the smallest.

The median price of the units available is \$70,000. To determine this, we order the prices from low (\$60,000) to high (\$275,000) and select the middle

value (\$70,000). For the median, the data must be at least an ordinal level of measurement.

Prices Ordered from Low to High		Prices Ordered from High to Low
\$ 60,000		\$275,000
65,000		80,000
70,000	← Median →	70,000
80,000		65,000
275,000		60,000

Median less affected by extreme values

Note that there is the same number of prices below the median of \$70,000 as above it. The median is, therefore, unaffected by extremely low or high prices. Had the highest price been \$90,000, or \$300,000, or even \$1 million, the median price would still be \$70,000. Likewise, had the lowest price been \$20,000 or \$50,000, the median price would still be \$70,000.

In the previous illustration there is an *odd* number of observations (five). How is the median determined for an *even* number of observations? As before, the observations are ordered. Then by convention to obtain a unique value we calculate the mean of the two middle observations. So for an even number of observations, the median may not be one of the given values.

Example

The three-year annualized total returns of the six top-performing diversified mutual funds are listed below. What is the median annualized return?

Solution



Name of Fund	Annualized Total Return
Artisan Mid Cap	42.10%
Clipper	15.50
Fidelity Advisor Mid-Cap	27.58
Fidelity Mid-Cap Stock	28.64
Smith Barney Aggressive	41.77
Van Kampen Comstock	16.97

Note that the number of returns is *even* (6). As before, first order the returns from low to high. Then identify the two middle returns. The arithmetic mean of the two middle observations gives us the median return. Arranging from low to high:

Clipper	15.50%	
Van Kampen Comstock	16.97	
Fidelity Advisor Mid-Cap	27.58	} 56.22/2 = 28.11 percent
Fidelity Mid-Cap Stock	28.64	
Smith Barney Aggressive	41.77	
Artisan Mid Cap	42.10	

Notice that the median is not one of the values. Also, half of the returns are below the median and half are above it.

Median can be determined for all levels of data except nominal

The major properties of the median are:

1. **It is not affected by extremely large or small values.** Therefore the median is a valuable measure of location when such values do occur.
2. **It can be computed for ordinal-level data or higher.** Recall from Chapter 1 that ordinal-level data can be ranked from low to high—such as the responses “excellent,” “very good,” “good,” “fair,” and “poor” to a question on a marketing survey. To use a simple illustration, suppose five people rated a new fudge bar. One person thought it was excellent, one rated it very good, one called it good, one rated it fair, and one considered it poor. The median response is “good.” Half of the responses are above “good”; the other half are below it.

The Mode

The **mode** is another measure of location.

MODE The value of the observation that appears most frequently.

The mode is especially useful in summarizing nominal-level data. As an example of its use for nominal-level data, a company has developed five bath oils. The bar chart in Chart 3–1 shows the results of a marketing survey designed to find which bath oil consumers prefer. The largest number of respondents favored Lamoure, as evidenced by the highest bar. Thus, Lamoure is the mode.

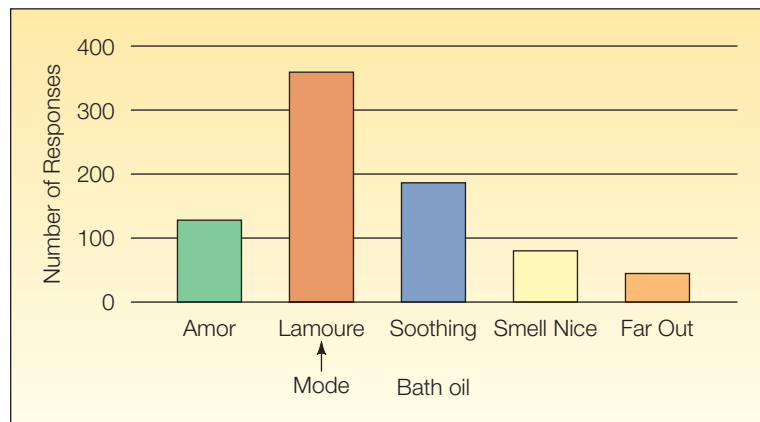


CHART 3–1 Number of Respondents Favoring Various Bath Oils

Example

The annual salaries of quality-control managers in selected states are shown below. What is the modal annual salary?

State	Salary	State	Salary	State	Salary
Arizona	\$35,000	Illinois	\$58,000	Ohio	\$50,000
California	49,100	Louisiana	60,000	Tennessee	60,000
Colorado	60,000	Maryland	60,000	Texas	71,400
Florida	60,000	Massachusetts	40,000	West Virginia	60,000
Idaho	40,000	New Jersey	65,000	Wyoming	55,000

Solution

A perusal of the salaries reveals that the annual salary of \$60,000 appears more often (six times) than any other salary. The mode is, therefore, \$60,000.

Disadvantages of the mode

In summary, we can determine the mode for all levels of data—nominal, ordinal, interval, and ratio. The mode also has the advantage of not being affected by extremely high or low values.

The mode does have disadvantages, however, that cause it to be used less frequently than the mean or median. For many sets of data, there is no mode because no value appears more than once. For example, there is no mode for this set of price data: \$19, \$21, \$23, \$20, and \$18. Since every value is different, however, it could be argued that every value is the mode. Conversely, for some data sets there is more than one mode. Suppose the ages of the individuals in a stock investment club are 22, 26, 27, 27, 31, 35, and 35. Both the ages 27 and 35 are modes. Thus, this grouping of ages is referred to as *bimodal* (having two modes). One would question the use of two modes to represent the location of this set of age data.

Self-Review 3–3



- A sample of single persons in Towson, Texas, receiving Social Security payments revealed these monthly benefits: \$852, \$598, \$580, \$1,374, \$960, \$878, and \$1,130.
 - What is the median monthly benefit?
 - How many observations are below the median? Above it?
- The number of work stoppages in the automobile industry for selected months are 6, 0, 10, 14, 8, and 0.
 - What is the median number of stoppages?
 - How many observations are below the median? Above it?
 - What is the modal number of work stoppages?

connect™

Exercises

- What would you report as the modal value for a set of observations if there were a total of:
 - 10 observations and no two values were the same?
 - 6 observations and they were all the same?
 - 6 observations and the values were 1, 2, 3, 3, 4, and 4?

For Exercises 18–20, determine the (a) mean, (b) median, and (c) mode.

- The following is the number of oil changes for the last 7 days at the Jiffy Lube located at the corner of Elm Street and Pennsylvania Avenue.

41	15	39	54	31	15	33
----	----	----	----	----	----	----

- The following is the percent change in net income from last year to this year for a sample of 12 construction companies in Denver.

5	1	-10	-6	5	12	7	8	2	5	-1	11
---	---	-----	----	---	----	---	---	---	---	----	----

- The following are the ages of the 10 people in the video arcade at the Southwyck Shopping Mall at 10 A.M.

12	8	17	6	11	14	8	17	10	8
----	---	----	---	----	----	---	----	----	---

- Several indicators of long-term economic growth in the United States are listed below.

Economic Indicator	Percent Change	Economic Indicator	Percent Change
Inflation	4.5%	Real GNP	2.9%
Exports	4.7	Investment (residential)	3.6
Imports	2.3	Investment (nonresidential)	2.1
Real disposable income	2.9	Productivity (total)	1.4
Consumption	2.7	Productivity (manufacturing)	5.2

- What is the median percent change?
- What is the modal percent change?

22. The total automobile sales (in millions of dollars) in the United States for the last 14 years are listed below. During this period, what was the median number of automobiles sold? What is the mode?

9.0	8.5	8.0	9.1	10.3	11.0	11.5	10.3	10.5	9.8	9.3	8.2	8.2	8.5
-----	-----	-----	-----	------	------	------	------	------	-----	-----	-----	-----	-----

23. The accounting firm of Rowatti and Koppel specializes in income tax returns for self-employed professionals, such as physicians, dentists, architects, and lawyers. The firm employs 11 accountants who prepare the returns. For last year the number of returns prepared by each accountant was:

58	75	31	58	46	65	60	71	45	58	80
----	----	----	----	----	----	----	----	----	----	----

Find the mean, median, and mode for the number of returns prepared by each accountant. If you could report only one, which measure of location would you recommend reporting?

24. The demand for the video games provided by Mid-Tech Video Games, Inc., has exploded in the last several years. Hence, the owner needs to hire several new technical people to keep up with the demand. Mid-Tech gives each applicant a special test that Dr. McGraw, the designer of the test, believes is closely related to the ability to create video games. For the general population the mean on this test is 100. Below are the scores on this test for the applicants.

95	105	120	81	90	115	99	100	130	10
----	-----	-----	----	----	-----	----	-----	-----	----

The president is interested in the overall quality of the job applicants based on this test. Compute the mean and the median score for the ten applicants. What would you report to the president? Does it seem that the applicants are better than the general population?

Software Solution

We can use a statistical software package to find many measures of location.

Example

Table 2–4 on page 28 shows the prices of the 80 vehicles sold last month at Whitner Autoplex in Raytown, Missouri. Determine the mean and the median selling price.

Solution

The mean and the median selling prices are reported in the following Excel output. (Remember: The instructions to create the output appear in the **Software Commands** section at the end of the chapter.) There are 80 vehicles in the study. So the calculations with a calculator would be tedious and prone to error.



Whitner Mean Median							
	A	B	C	D	E	F	G
1	Price	Price(\$000)	Age	Type		<i>Price</i>	
2	23197	23.197	46	0			
3	23372	23.372	48	0		Mean	23218.1625
4	20454	20.454	40	1		Standard Error	486.8409474
5	23591	23.591	40	0		Median	22831
6	26651	26.651	46	1		Mode	20642
7	27453	27.453	37	1		Standard Deviation	4354.43781
8	17266	17.266	32	1		Sample Variance	18961128.64
9	18021	18.021	29	1		Kurtosis	0.5433087
10	28683	28.683	38	1		Skewness	0.72681585
11	30872	30.872	43	0		Range	20379
12	19587	19.587	32	0		Minimum	15546
13	23169	23.169	47	0		Maximum	35925
14	35851	35.851	56	0		Sum	1857453
15	19251	19.251	42	1		Count	80
16	20047	20.047	28	1			

The mean selling price is \$23,218 and the median is \$22,831. These two values are less than \$400 apart. So either value is reasonable. We can also see from the Excel output that there were 80 vehicles sold and their total price is \$1,857,453. We will describe the meaning of standard error, standard deviation, and other measures later.

What can we conclude? The typical vehicle sold for about \$23,000. Ms. Ball of AutoUSA might use this value in her revenue projections. For example, if the dealership could increase the number sold in a month from 80 to 90, this would result in an additional estimated \$230,000 of revenue, found by $10 \times \$23,000$.

The Relative Positions of the Mean, Median, and Mode

For a symmetric, mound-shaped distribution, mean, median, and mode are equal.

Refer to the histogram in Chart 3–2. It is a symmetric distribution, which is also mound-shaped. This distribution *has the same shape on either side of the center*. If the polygon were folded in half, the two halves would be identical. For any symmetric distribution, the mode, median, and mean are located at the center and are always equal. They are all equal to 20 years in Chart 3–2. We should point out that there are symmetric distributions that are not mound-shaped.

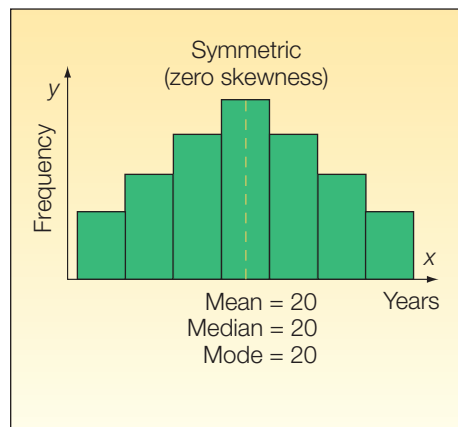


CHART 3–2 A Symmetric Distribution

The number of years corresponding to the highest point of the curve is the *mode* (20 years). Because the distribution is symmetrical, the *median* corresponds to the point where the distribution is cut in half (20 years). The total number of frequencies representing many years is offset by the total number representing few years, resulting in an *arithmetic mean* of 20 years. Logically, any of the three measures would be appropriate to represent the distribution's center.

A skewed distribution is not symmetrical.

If a distribution is nonsymmetrical, or **skewed**, the relationship among the three measures changes. In a **positively skewed distribution**, the arithmetic mean is the largest of the three measures. Why? Because the mean is influenced more than the median or mode by a few extremely high values. The median is generally the next largest measure in a positively skewed frequency distribution. The mode is the smallest of the three measures.

If the distribution is highly skewed, such as the weekly incomes in Chart 3–3, the mean would not be a good measure to use. The median and mode would be more representative.

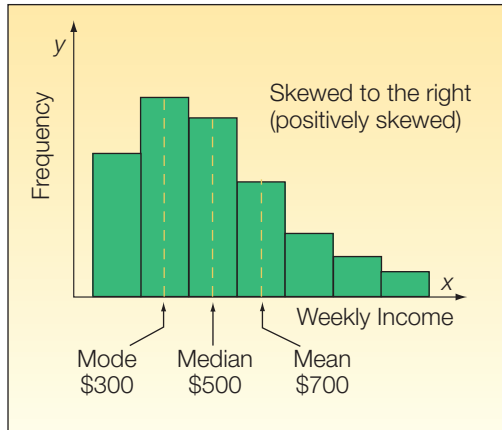


CHART 3-3 A Positively Skewed Distribution

Conversely, if a distribution is **negatively skewed**, the mean is the lowest of the three measures. The mean is, of course, influenced by a few extremely low observations. The median is greater than the arithmetic mean, and the modal value is the largest of the three measures. Again, if the distribution is highly skewed, such as the distribution of tensile strengths shown in Chart 3-4, the mean should not be used to represent the data.

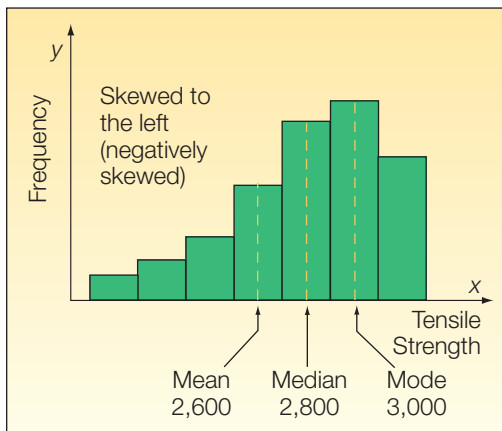


CHART 3-4 A Negatively Skewed Distribution

Self-Review 3-4



The weekly sales from a sample of Hi-Tec electronic supply stores were organized into a frequency distribution. The mean of weekly sales was computed to be \$105,900, the median \$105,000, and the mode \$104,500.

- Sketch the sales in the form of a smoothed frequency polygon. Note the location of the mean, median, and mode on the X-axis.
- Is the distribution symmetrical, positively skewed, or negatively skewed? Explain.



Exercises

25. The unemployment rate in the state of Alaska by month is given in the table below:

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
8.7	8.8	8.7	7.8	7.3	7.8	6.6	6.5	6.5	6.8	7.3	7.6

- What is the arithmetic mean of the Alaska unemployment rates?
 - Find the median and the mode for the unemployment rates.
 - Compute the arithmetic mean and median for just the winter (Dec–Mar) months. Is it much different?
26. Big Orange Trucking is designing an information system for use in “in-cab” communications. It must summarize data from eight sites throughout a region to describe typical conditions. Compute an appropriate measure of central location for each of the three variables shown in the table below:

City	Wind Direction	Temperature	Pavement
Anniston, AL	West	89	Dry
Atlanta, GA	Northwest	86	Wet
Augusta, GA	Southwest	92	Wet
Birmingham, AL	South	91	Dry
Jackson, MS	Southwest	92	Dry
Meridian, MS	South	92	Trace
Monroe, LA	Southwest	93	Wet
Tuscaloosa, AL	Southwest	93	Trace

The Geometric Mean

The geometric mean is never greater than the arithmetic mean.

The geometric mean is useful in finding the average change of percentages, ratios, indexes, or growth rates over time. It has a wide application in business and economics because we are often interested in finding the percentage changes in sales, salaries, or economic figures, such as the Gross Domestic Product, which compound or build on each other. The geometric mean of a set of n positive numbers is defined as the n th root of the product of n values. The formula for the geometric mean is written:

GEOMETRIC MEAN

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

[3–4]

The geometric mean will always be less than or equal to (never more than) the arithmetic mean. Also all the data values must be positive.

As an example of the geometric mean, suppose you receive a 5 percent increase in salary this year and a 15 percent increase next year. The average annual percent increase is 9.886, not 10.0. Why is this so? We begin by calculating the geometric mean. Recall, for example, that a 5 percent increase in salary is 105 percent. We will write it as 1.05.

$$GM = \sqrt{(1.05)(1.15)} = 1.09886$$

This can be verified by assuming that your monthly earning was \$3,000 to start and you received two increases of 5 percent and 15 percent.

$$\text{Raise 1} = \$3,000 (.05) = \$150.00$$

$$\text{Raise 2} = \$3,150 (.15) = \underline{472.50}$$

$$\text{Total} \qquad \qquad \qquad \underline{\$622.50}$$

Your total salary increase is \$622.50. This is equivalent to:

$$\$3,000.00 (.09886) = \$296.58$$

$$\$3,296.58 (.09886) = \frac{325.90}{\$622.48 \text{ is about } \$622.50}$$

The following example shows the geometric mean of several percentages.

Example

The return on investment earned by Atkins Construction Company for four successive years was: 30 percent, 20 percent, -40 percent, and 200 percent. What is the geometric mean rate of return on investment?

Solution

The number 1.3 represents the 30 percent return on investment, which is the “original” investment of 1.0 plus the “return” of 0.3. The number 0.6 represents the loss of 40 percent, which is the original investment of 1.0 less the loss of 0.4. This calculation assumes the total return each period is reinvested or becomes the base for the next period. In other words, the base for the second period is 1.3 and the base for the third period is (1.3)(1.2) and so forth.

Then the geometric mean rate of return is 29.4 percent, found by

$$GM = \sqrt[n]{(X_1)(X_2) \cdots (X_n)} = \sqrt[4]{(1.3)(1.2)(0.6)(3.0)} = \sqrt[4]{2.808} = 1.294$$

The geometric mean is the fourth root of 2.808. So, the average rate of return (compound annual growth rate) is 29.4 percent.

Notice also that if you compute the arithmetic mean $[(30 + 20 - 40 + 200)/4 = 52.5]$, you would have a much larger number, which would overstate the true rate of return!

A second application of the geometric mean is to find an average percent change over a period of time. For example, if you earned \$30,000 in 1997 and \$50,000 in 2007, what is your annual rate of increase over the period? It is 5.24 percent. The rate of increase is determined from the following formula.

**RATE OF INCREASE
OVER TIME**

$$GM = \sqrt[n]{\frac{\text{Value at end of period}}{\text{Value at start of period}}} - 1 \quad [3-5]$$

In the above box n is the number of periods. An example will show the details of finding the average annual percent increase.

Example

During the decade of the 1990s, and into the 2000s, Las Vegas, Nevada, was the fastest-growing city in the United States. The population increased from 258,295 in 1990 to 552,539 in 2007. This is an increase of 294,244 people or a 113.9 percent increase over the 17-year period. It has more than doubled over the period. What is the average *annual* percent increase?

Solution

There are 17 years between 1990 and 2007 so $n = 17$. Then formula (3-5) for the geometric mean as applied to this problem is:

$$GM = \sqrt[n]{\frac{\text{Value at end of period}}{\text{Value at start of period}}} - 1.0 = \sqrt[17]{\frac{552,539}{258,295}} - 1.0 = 1.0457 - 1.0 = .0457$$

The value of .0457 indicates that the average annual growth over the 17-year period was 4.57 percent. To put it another way, the population of Las Vegas increased at a rate of 4.57 percent per year from 1990 to 2007.

Self-Review 3–5



- The percent increase in sales for the last 4 years at Combs Cosmetics were: 4.91, 5.75, 8.12, and 21.60.
 - Find the geometric mean percent increase.
 - Find the arithmetic mean percent increase.
 - Is the arithmetic mean equal to or greater than the geometric mean?
- Production of Cablos trucks increased from 23,000 units in 1998 to 120,520 units in 2008. Find the geometric mean annual percent increase.

Exercises

connect™

- Compute the geometric mean of the following percent increases: 8, 12, 14, 26, and 5.
- Compute the geometric mean of the following percent increases: 2, 8, 6, 4, 10, 6, 8, and 4.
- Listed below is the percent increase in sales for the MG Corporation over the last 5 years. Determine the geometric mean percent increase in sales over the period.

9.4	13.8	11.7	11.9	14.7
-----	------	------	------	------

- In 1996 a total of 14,968,000 taxpayers in the United States filed their individual tax returns electronically. By the year 2005 the number increased to 68,476,000. What is the geometric mean annual increase for the period?
- The Consumer Price Index is reported monthly by the U.S. Bureau of Labor Statistics. It reports the change in prices for a market basket of goods from one period to another. The index for 1994 was 148.2, by 2007 it increased to 210.2. What was the geometric mean annual increase for the period?
- In 1976 the nationwide average price of a gallon of unleaded gasoline at a self-serve pump was \$0.605. By 2008 the average price had increased to \$3.23. What was the geometric mean annual increase for the period?
- In 1985 there were 340,213 cell phone subscribers in the United States. By 2006 the number of cell phone subscribers increased to 233,000,000. What is the geometric mean annual increase for the period?
- The information below shows the cost for a year of college in public and private colleges in 1992 and 2007. What is the geometric mean annual increase for the period for the two types of colleges? Compare the rates of increase.

Type of College	1992	2007
Public	\$ 4,975	\$ 11,354
Private	12,284	27,516



Statistics in Action

The United States Postal Service has tried to become more “user friendly” in the last several years. A recent survey showed that customers were interested in more *consistency* in the time it takes to make a delivery. Under the old conditions, a local letter might take only one day to deliver, or it might take several. “Just tell me how many days ahead I need to mail the birthday card to Mom so it gets there on her birthday, not early, not late,” was a common complaint. The level of consistency is measured by the standard deviation of the delivery times.

Why Study Dispersion?

A measure of location, such as the mean or the median, only describes the center of the data. It is valuable from that standpoint, but it does not tell us anything about the spread of the data. For example, if your nature guide told you that the river ahead averaged 3 feet in depth, would you want to wade across on foot without additional information? Probably not. You would want to know something about the variation in the depth. Is the maximum depth of the river 3.25 feet and the minimum 2.75 feet? If that is the case, you would probably agree to cross. What if you learned the river depth ranged from 0.50 feet to 5.5 feet? Your decision would probably be not to cross. Before making a decision about crossing the river, you want information on both the typical depth and the dispersion in the depth of the river.

A small value for a measure of dispersion indicates that the data are clustered closely, say, around the arithmetic mean. The mean is therefore considered representative of the data. Conversely, a large measure of dispersion indicates that the mean is not reliable. Refer to Chart 3–5. The 100 employees of Hammond Iron

Works, Inc., a steel fabricating company, are organized into a histogram based on the number of years of employment with the company. The mean is 4.9 years, but the spread of the data is from 6 months to 16.8 years. The mean of 4.9 years is not very representative of all the employees.

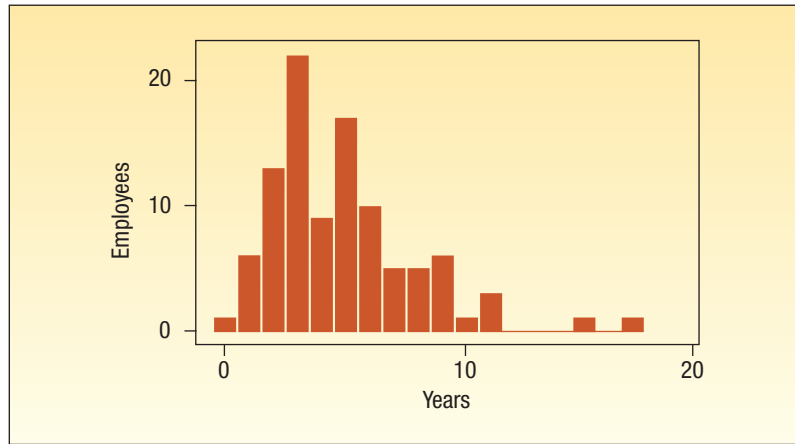


CHART 3-5 Histogram of Years of Employment at Hammond Iron Works, Inc.

The average is not representative because of the large spread.

A second reason for studying the dispersion in a set of data is to compare the spread in two or more distributions. Suppose, for example, that the new Vision Quest LCD computer monitor is assembled in Baton Rouge and also in Tucson. The arithmetic mean hourly output in both the Baton Rouge plant and the Tucson plant is 50. Based on the two means, you might conclude that the distributions of the hourly outputs are identical. Production records for 9 hours at the two plants, however, reveal that this conclusion is not correct (see Chart 3-6). Baton Rouge production varies from 48 to 52 assemblies per hour. Production at the Tucson plant is more erratic, ranging from 40 to 60 per hour. Therefore, the hourly output for Baton Rouge is clustered near the mean of 50; the hourly output for Tucson is more dispersed.

A measure of dispersion can be used to evaluate the reliability of two or more measures of location.

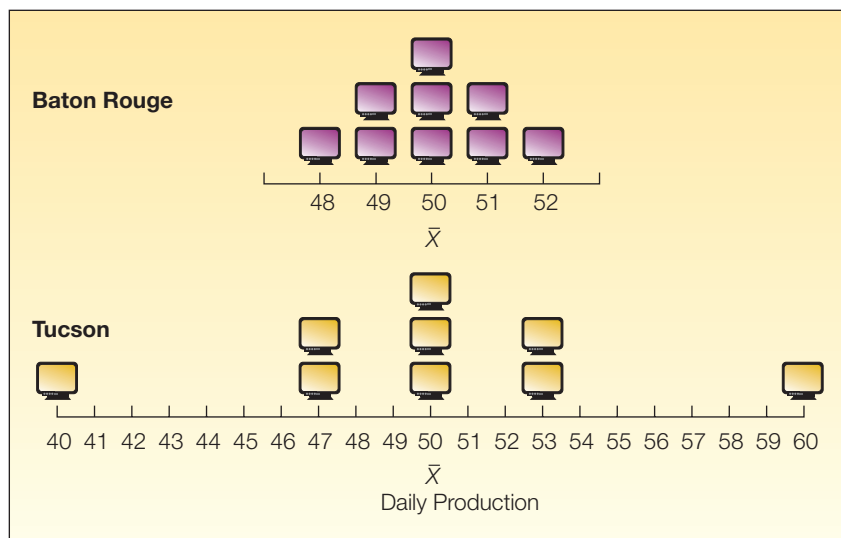


CHART 3-6 Hourly Production of Computer Monitors at the Baton Rouge and Tucson Plants

Measures of Dispersion

We will consider several measures of dispersion. The range is based on the largest and the smallest values in the data set, that is, only two values are considered. The mean deviation, the variance, and the standard deviation use all the values in a data set and are all based on deviations from the arithmetic mean.

Range

The simplest measure of dispersion is the **range**. It is the difference between the largest and the smallest values in a data set. In the form of an equation:

RANGE

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

[3-6]

The range is widely used in statistical process control (SPC) applications because it is very easy to calculate and understand.

Example

Refer to Chart 3-6 on the previous page. Find the range in the number of computer monitors produced per hour for the Baton Rouge and the Tucson plants. Interpret the two ranges.

Solution

The range of the hourly production of computer monitors at the Baton Rouge plant is 4, found by the difference between the largest hourly production of 52 and the smallest of 48. The range in the hourly production for the Tucson plant is 20 computer monitors, found by $60 - 40$. We therefore conclude that (1) there is less dispersion in the hourly production in the Baton Rouge plant than in the Tucson plant because the range of 4 computer monitors is less than a range of 20 computer monitors, and (2) the production is clustered more closely around the mean of 50 at the Baton Rouge plant than at the Tucson plant (because a range of 4 is less than a range of 20). Thus, the mean production in the Baton Rouge plant (50 computer monitors) is a more representative measure of location than the mean of 50 computer monitors for the Tucson plant.

Mean Deviation

A defect of the range is that it is based on only two values, the highest and the lowest; it does not take into consideration all of the values. The **mean deviation** does. It measures the mean amount by which the values in a population, or sample, vary from their mean. In terms of a definition:

MEAN DEVIATION The arithmetic mean of the absolute values of the deviations from the arithmetic mean.

In terms of a formula, the mean deviation, designated MD , is computed for a sample by:

MEAN DEVIATION

$$MD = \frac{\sum |X - \bar{X}|}{n}$$

[3-7]

where:

X is the value of each observation.

\bar{X} is the arithmetic mean of the values.

n is the number of observations in the sample.

$||$ indicates the absolute value.

Why do we ignore the signs of the deviations from the mean? If we didn't, the positive and negative deviations from the mean would exactly offset each other, and the mean deviation would always be zero. Such a measure (zero) would be a useless statistic.

Example



Solution

The number of cappuccinos sold at the Starbucks location in the Orange County Airport between 4 and 7 p.m. for a sample of 5 days last year were 20, 40, 50, 60, and 80. In the LAX airport in Los Angeles, the number of cappuccinos sold at a Starbucks location between 4 and 7 p.m. for a sample of 5 days last year were 20, 49, 50, 51, and 80. Determine the mean, median, range, and mean deviation for each location. Compare the differences.

For the Orange County location the mean, median, and range are:

Mean	50 cappuccinos per day
Median	50 cappuccinos per day
Range	60 cappuccinos per day

The mean deviation is the mean of the absolute differences between individual observations and the arithmetic mean. For Orange County, the mean number of cappuccinos sold is 50, found by $(20 + 40 + 50 + 60 + 80)/5$. Next we find the differences between each observation and the mean. Then we sum these differences, ignoring the signs, and divide the sum by the number of observations. The result is the mean difference between the observations and the mean.

Number of Cappuccinos Sold Daily	$(X - \bar{X})$	Absolute Deviation
20	$(20 - 50) = -30$	30
40	$(40 - 50) = -10$	10
50	$(50 - 50) = 0$	0
60	$(60 - 50) = 10$	10
80	$(80 - 50) = 30$	30
		Total 80

$$MD = \frac{\sum |X - \bar{X}|}{n} = \frac{80}{5} = 16$$

The mean deviation is 16 cappuccinos per day and shows that the number of cappuccinos sold deviates, on average, by 16 from the mean of 50 cappuccinos per day.

The summary of the mean, median, range, and mean deviation for LAX follows. You should perform the calculations to verify the results.

Mean	50 cappuccinos per day
Median	50 cappuccinos per day
Range	60 cappuccinos per day
Mean Deviation	12.4 cappuccinos per day

Recall in the previous chapter that we described data using graphical methods. In this chapter, we describe data using numerical measures. When we use numerical measures, it is very important to always report measures of location and dispersion.

Let's interpret and compare the results of our measures for the Starbucks locations. The mean and median of the two locations are exactly the same, 50 cappuccinos per day. Therefore, the location of both distributions is the same. The range for both locations is also the same, 60. However, recall that the range provides limited information about the dispersion of the distribution.

Notice that the mean deviations are not the same because they are based on the differences between all observations and the arithmetic mean, which show the relative closeness or clustering of the data relative to the mean or center of the distribution. Compare the mean deviation for Orange County of 16 to the mean deviation for LAX of 12.4. Based on the mean deviation, we can say that the dispersion for the sales distribution of the LAX Starbucks is more concentrated near the mean of 50 than the Orange County location.

Advantages of mean deviation

The mean deviation has two advantages. First, it uses all the values in the computation. Recall that the range uses only the highest and the lowest values. Second, it is easy to understand—it is the average amount by which values deviate from the mean. However, its drawback is the use of absolute values. Generally, absolute values are difficult to work with and to explain, so the mean deviation is not used as frequently as other measures of dispersion, such as the standard deviation.

Self-Review 3–6

The weights of containers being shipped to Ireland are (in thousands of pounds):



95 103 105 110 104 105 112 90

- What is the range of the weights?
- Compute the arithmetic mean weight.
- Compute the mean deviation of the weights.

connect™

Exercises

For Exercises 35–38, calculate the (a) range, (b) arithmetic mean, (c) mean deviation and (d) the range. Interpret your values.

- There were five customer service representatives on duty at the Electronic Super Store during last weekend's sale. The numbers of HDTVs these representatives sold are: 5, 8, 4, 10, and 3.
- The Department of Statistics at Western State University offers eight sections of basic statistics. Following are the numbers of students enrolled in these sections: 34, 46, 52, 29, 41, 38, 36, and 28.
- Dave's Automatic Door installs automatic garage door openers. The following list indicates the number of minutes needed to install a sample of 10 door openers: 28, 32, 24, 46, 44, 40, 54, 38, 32, and 42.

38. A sample of eight companies in the aerospace industry was surveyed as to their return on investment last year. The results are (in percent): 10.6, 12.6, 14.8, 18.2, 12.0, 14.8, 12.2, and 15.6.
39. Ten randomly selected young adults living in California rated the taste of a newly developed sushi pizza topped with tuna, rice, and kelp on a scale of 1 to 50, with 1 indicating they did not like the taste and 50 that they did. The ratings were:

34	39	40	46	33	31	34	14	15	45
----	----	----	----	----	----	----	----	----	----

In a parallel study 10 randomly selected young adults in Iowa rated the taste of the same pizza. The ratings were:

28	25	35	16	25	29	24	26	17	20
----	----	----	----	----	----	----	----	----	----

As a market researcher, compare the potential markets for sushi pizza.

40. A sample of the personnel files of eight employees at the Pawnee location of Acme Carpet Cleaners, Inc., revealed that during the last six-month period they lost the following number of days due to illness:

2	0	6	3	10	4	1	2
---	---	---	---	----	---	---	---

A sample of eight employees during the same period at the Chickpea location of Acme Carpets revealed they lost the following number of days due to illness.

2	0	1	0	5	0	1	0
---	---	---	---	---	---	---	---

As the director of human relations, compare the two locations. What would you recommend?

Variance and Standard Deviation

Variance and standard deviation are based on squared deviations from the mean.

The **variance** and **standard deviation** are also based on the deviations from the mean. However, instead of using the absolute value of the deviations, the variance and the standard deviation square the deviations.

VARIANCE The arithmetic mean of the squared deviations from the mean.

The variance is nonnegative and is zero only if all observations are the same.

STANDARD DEVIATION The square root of the variance.

Population Variance The formulas for the population variance and the sample variance are slightly different. The population variance is considered first. (Recall that a population is the totality of all observations being studied.) The **population variance** is found by:

POPULATION VARIANCE

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

[3-8]

Where:

σ^2 is the population variance (σ is the lowercase Greek letter sigma). It is read as “sigma squared.”

X is the value of an observation in the population.

μ is the arithmetic mean of the population.

N is the number of observations in the population.

Note the process of computing the variance.

- We begin by finding the mean.
- Next we find the difference between each observation and the mean, and square that difference.
- Then we sum all the squared differences.
- And finally we divide the sum of the squared differences by the number of items in the population.

So you might think of the population variance as the mean of the squared difference between each value and the mean. For populations whose values are near the mean, the variance will be small. For populations whose values are dispersed from the mean, the population variance will be large.

The variance overcomes the weakness of the range by using all the values in the population, whereas the range uses only the largest and the smallest. We overcome the issue where $\sum(X - \mu) = 0$ by squaring the differences, instead of using the absolute values. Squaring the differences will always result in non-negative values.

Example

The number of traffic citations issued last year by month in Beaufort County, South Carolina, is reported below.

Month	January	February	March	April	May	June	July	August	September	October	November	December
Citations	19	17	22	18	28	34	45	39	38	44	34	10

Determine the population variance.

Solution

Because we are studying all the citations for a year, the data is a population. To determine the population variance, we use formula [3-8]. The table below details the calculations.

Month	Citations (X)	$X - \mu$	$(X - \mu)^2$
January	19	-10	100
February	17	-12	144
March	22	-7	49
April	18	-11	121
May	28	-1	1
June	34	5	25
July	45	16	256
August	39	10	100
September	38	9	81
October	44	15	225
November	34	5	25
December	10	-19	361
Total	348	0	1,488

1. We begin by determining the arithmetic mean of the population. The total number of citations issued for the year is 348, so the mean number issued per month is 29.

$$\mu = \frac{\sum X}{N} = \frac{19 + 17 + \cdots + 10}{12} = \frac{348}{12} = 29$$

2. Next we find the difference between each observation and the mean. This is shown in the third column of the table. Recall that earlier in the chapter (page 59) we indicated that the sum of the differences between each value and the mean is 0. From the spreadsheet, the sum of the differences between the mean and the number of citations each month is 0.
3. The next step is to square the difference between each monthly value. That is shown in the fourth column of the table. By squaring the differences, we convert both the positive and the negative values to a plus sign. Hence, each difference will be positive.
4. The squared differences are totaled. The total of the fourth column is 1,488. That is the term $\sum(X - \mu)^2$.
5. Finally, we divide the squared differences by N , the number of observations in the population.

$$\sigma^2 = \frac{\sum(X - \mu)^2}{N} = \frac{1488}{12} = 124$$

So, the population variation for the number of citations is 124.

Like the range and the mean deviation, the variance can be used to compare dispersion in two or more sets of observations. For example, the variance for the number of citations issued in Beaufort County was just computed to be 124. If the variance in the number of citations issued in Marlboro County, South Carolina, is 342.9, we conclude that (1) there is less dispersion in the distribution of the number of citations issued in Beaufort County than in Marlboro County (because 124 is less than 342.9); and (2) the number of citations in Beaufort County is more closely clustered around the mean of 29 than for the number of citations issued in Marlboro County. Thus the mean number of citations issued in Beaufort County is a more representative measure of location than the mean number of citations in Marlboro County.

Variance is difficult to interpret because the units are squared.

Standard deviation is in the same units as the data.

Population Standard Deviation Both the range and the mean deviation are easy to interpret. The range is the difference between the high and low values of a set of data, and the mean deviation is the mean of the deviations from the mean. However, the variance is difficult to interpret for a single set of observations. The variance of 124 for the number of citations issued is not in terms of citations, but citations squared.

There is a way out of this difficulty. By taking the square root of the population variance, we can transform it to the same unit of measurement used for the original data. The square root of 124 citations-squared is 11.14 citations. The units are now simply citations. The square root of the population variance is the **population standard deviation**.

POPULATION STANDARD DEVIATION

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

[3-9]

Self-Review 3–7

The Philadelphia office of Price Waterhouse Coopers LLP hired five accounting trainees this year. Their monthly starting salaries were: \$3,536; \$3,173; \$3,448; \$3,121; and \$3,622.

- Compute the population mean.
- Compute the population variance.
- Compute the population standard deviation.
- The Pittsburgh office hired six trainees. Their mean monthly salary was \$3,550, and the standard deviation was \$250. Compare the two groups.

connect™

Exercises

- Consider these five values a population: 8, 3, 7, 3, and 4.
 - Determine the mean of the population.
 - Determine the variance.
- Consider these six values a population: 13, 3, 8, 10, 8, and 6.
 - Determine the mean of the population.
 - Determine the variance.
- The annual report of Dennis Industries cited these primary earnings per common share for the past 5 years: \$2.68, \$1.03, \$2.26, \$4.30, and \$3.58. If we assume these are population values, what is:
 - The arithmetic mean primary earnings per share of common stock?
 - The variance?
- Referring to Exercise 43, the annual report of Dennis Industries also gave these returns on stockholder equity for the same five-year period (in percent): 13.2, 5.0, 10.2, 17.5, and 12.9.
 - What is the arithmetic mean return?
 - What is the variance?
- Plywood, Inc., reported these returns on stockholder equity for the past 5 years: 4.3, 4.9, 7.2, 6.7, and 11.6. Consider these as population values.
 - Compute the range, the arithmetic mean, the variance, and the standard deviation.
 - Compare the return on stockholder equity for Plywood, Inc., with that for Dennis Industries cited in Exercise 44.
- The annual incomes of the five vice presidents of TMV Industries are: \$125,000; \$128,000; \$122,000; \$133,000; and \$140,000. Consider this a population.
 - What is the range?
 - What is the arithmetic mean income?
 - What is the population variance? The standard deviation?
 - The annual incomes of officers of another firm similar to TMV Industries were also studied. The mean was \$129,000 and the standard deviation \$8,612. Compare the means and dispersions in the two firms.

Sample Variance The formula for the population mean is $\mu = \Sigma X/N$. We just changed the symbols for the sample mean; that is, $\bar{X} = \Sigma X/n$. Unfortunately, the conversion from the population variance to the sample variance is not as direct. It requires a change in the denominator. Instead of substituting n (number in the sample) for N (number in the population), the denominator is $n - 1$. Thus the formula for the **sample variance** is:

SAMPLE VARIANCE

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

[3–10]

where:

s^2 is the sample variance.

X is the value of each observation in the sample.

\bar{X} is the mean of the sample.

n is the number of observations in the sample.

Why is this change made in the denominator? Although the use of n is logical since \bar{X} is used to estimate μ , it tends to underestimate the population variance, σ^2 . The use of $(n - 1)$ in the denominator provides the appropriate correction for this tendency. Because the primary use of sample statistics like s^2 is to estimate population parameters like σ^2 , $(n - 1)$ is preferred to n in defining the sample variance. We will also use this convention when computing the sample standard deviation.

Example

The hourly wages for a sample of part-time employees at Home Depot are: \$12, \$20, \$16, \$18, and \$19. What is the sample variance?

Solution

The sample variance is computed by using formula (3-10).

$$\bar{X} = \frac{\sum X}{n} = \frac{\$85}{5} = \$17$$

Hourly Wage (X)	$X - \bar{X}$	$(X - \bar{X})^2$
\$12	-\$5	25
20	3	9
16	-1	1
18	1	1
19	2	4
<u>\$85</u>	<u>0</u>	<u>40</u>

$$s^2 = \frac{\sum(X - \bar{X})^2}{n - 1} = \frac{40}{5 - 1}$$

$$= 10 \text{ in dollars squared}$$

Sample Standard Deviation The sample standard deviation is used as an estimator of the population standard deviation. As noted previously, the population standard deviation is the square root of the population variance. Likewise, the *sample standard deviation is the square root of the sample variance*. The sample standard deviation is most easily determined by:

SAMPLE STANDARD DEVIATION

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

[3-11]

Example

The sample variance in the previous example involving hourly wages was computed to be 10. What is the sample standard deviation?

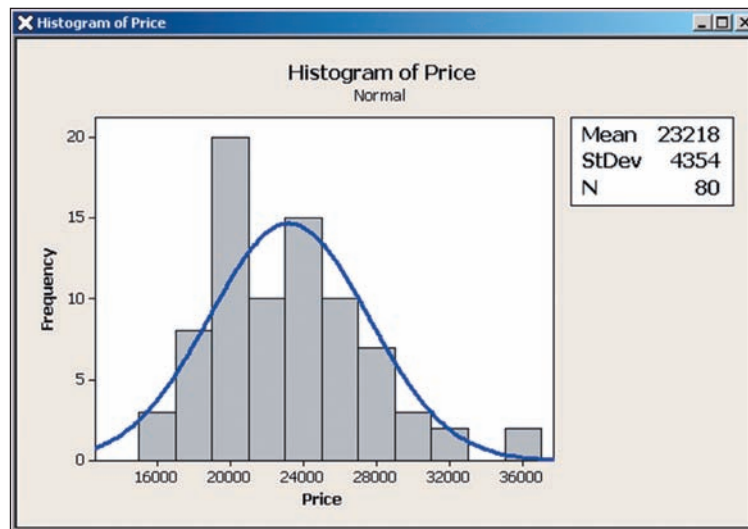
Solution

The sample standard deviation is \$3.16, found by $\sqrt{10}$. Note again that the sample variance is in terms of dollars squared, but taking the square root of 10 gives us \$3.16, which is in the same units (dollars) as the original data.

Software Solution

On page 66 we used Excel to determine the mean and median of the Whitner Autoplex sales data. You will also note that it outputs the sample standard deviation. Excel, like most other statistical software, assumes the data are from a sample.

Another software package that we will use in this text is MINITAB. This package uses a spreadsheet format, much like Excel, but produces a wider variety of statistical output. The information for the Whitner Autoplex selling prices follows. Note that a histogram (although the default is to use a class interval of \$2,000 and 11 classes) is included as well as the mean, sample standard deviation and the number of observations. A graph of the normal curve is superimposed on the frequency distribution. We will explain the normal curve in Chapter 7.



Self-Review 3–8



The years of service for a sample of seven employees at a State Farm Insurance claims office in Cleveland, Ohio, are: 4, 2, 5, 4, 5, 2, and 6. What is the sample variance? Compute the sample standard deviation.

Exercises

For Exercises 47–52, do the following:

- a. Compute the sample variance.
 - b. Determine the sample standard deviation.
47. Consider these values a sample: 7, 2, 6, 2, and 3.
 48. The following five values are a sample: 11, 6, 10, 6, and 7.



Statistics in Action

Most colleges report the “average class size.” This information can be misleading because average class size can be found several ways. If we find the number of students in *each class* at a particular university, the result is the mean number of students per class. If we compile a list of the class sizes for each student and find the mean class size, we might find the mean to be quite different. One school found the mean number of students in each of its 747 classes to be 40. But when it found the mean from a list of the class sizes of each student it was 147. Why the disparity? Because there are few students in the small classes and a larger number of students in the larger classes, which has the effect of increasing the mean class size when it is calculated this way. A school could reduce this mean class size for each student by reducing the number of students in each class. That is, cut out the large freshman lecture classes.

49. Dave's Automatic Door, referred to in Exercise 37, installs automatic garage door openers. Based on a sample, following are the times, in minutes, required to install 10 door openers: 28, 32, 24, 46, 44, 40, 54, 38, 32, and 42.
50. The sample of eight companies in the aerospace industry, referred to in Exercise 36, was surveyed as to their return on investment last year. The results are: 10.6, 12.6, 14.8, 18.2, 12.0, 14.8, 12.2, and 15.6.
51. The Houston, Texas, Motel Owner Association conducted a survey regarding weekday motel rates in the area. Listed below is the room rate for business-class guests for a sample of 10 motels.

\$101	\$97	\$103	\$110	\$78	\$87	\$101	\$80	\$106	\$88
-------	------	-------	-------	------	------	-------	------	-------	------

52. A consumer watchdog organization is concerned about credit card debt. A survey of 10 young adults with credit card debt of more than \$2,000 showed they paid an average of just over \$100 per month against their balances. Listed below is the amounts each young adult paid last month.

\$110	\$126	\$103	\$93	\$99	\$113	\$87	\$101	\$109	\$100
-------	-------	-------	------	------	-------	------	-------	-------	-------

Interpretation and Uses of the Standard Deviation

The standard deviation is commonly used as a measure to compare the spread in two or more sets of observations. For example, the standard deviation of the biweekly amounts invested in the Dupree Paint Company profit-sharing plan is computed to be \$7.51. Suppose these employees are located in Georgia. If the standard deviation for a group of employees in Texas is \$10.47, and the means are about the same, it indicates that the amounts invested by the Georgia employees are not dispersed as much as those in Texas (because $\$7.51 < \10.47). Since the amounts invested by the Georgia employees are clustered more closely about the mean, the mean for the Georgia employees is a more reliable measure than the mean for the Texas group.

Chebyshev's Theorem

We have stressed that a small standard deviation for a set of values indicates that these values are located close to the mean. Conversely, a large standard deviation reveals that the observations are widely scattered about the mean. The Russian mathematician P. L. Chebyshev (1821–1894) developed a theorem that allows us to determine the minimum proportion of the values that lie within a specified number of standard deviations of the mean. For example, according to **Chebyshev's theorem**, at least three of four values, or 75 percent, must lie between the mean plus two standard deviations and the mean minus two standard deviations. This relationship applies regardless of the shape of the distribution. Further, at least eight of nine values, or 88.9 percent, will lie between plus three standard deviations and minus three standard deviations of the mean. At least 24 of 25 values, or 96 percent, will lie between plus and minus five standard deviations of the mean.

Chebyshev's theorem states:

CHEBYSHEV'S THEOREM For any set of observations (sample or population), the proportion of the values that lie within k standard deviations of the mean is at least $1 - 1/k^2$, where k is any constant greater than 1.

Example

The arithmetic mean biweekly amount contributed by the Dupree Paint employees to the company's profit-sharing plan is \$51.54, and the standard deviation is \$7.51. At least what percent of the contributions lie within plus 3.5 standard deviations and minus 3.5 standard deviations of the mean?

Solution

About 92 percent, found by

$$1 - \frac{1}{k^2} = 1 - \frac{1}{(3.5)^2} = 1 - \frac{1}{12.25} = 0.92$$

The Empirical Rule applies only to symmetrical, bell-shaped distributions.

The Empirical Rule

Chebyshev's theorem is concerned with any set of values; that is, the distribution of values can have any shape. However, for a symmetrical, bell-shaped distribution such as the one in Chart 3-7, we can be more precise in explaining the dispersion about the mean. These relationships involving the standard deviation and the mean are described by the **Empirical Rule**, sometimes called the **Normal Rule**.

EMPIRICAL RULE For a symmetrical, bell-shaped frequency distribution, approximately 68 percent of the observations will lie within plus and minus one standard deviation of the mean; about 95 percent of the observations will lie within plus and minus two standard deviations of the mean; and practically all (99.7 percent) will lie within plus and minus three standard deviations of the mean.

These relationships are portrayed graphically in Chart 3-7 for a bell-shaped distribution with a mean of 100 and a standard deviation of 10.

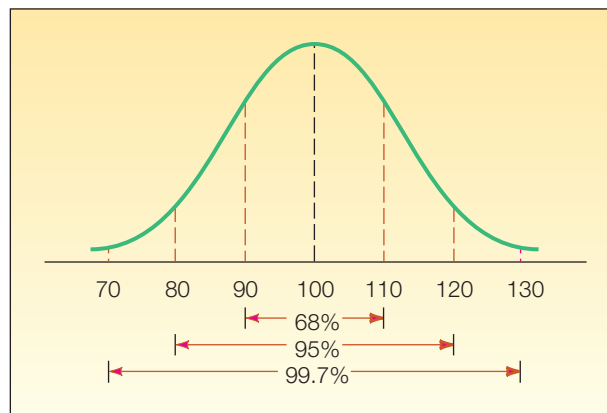


CHART 3-7 A Symmetrical, Bell-Shaped Curve Showing the Relationships between the Standard Deviation and the Observations

It has been noted that if a distribution is symmetrical and bell-shaped, practically all of the observations lie between the mean plus and minus three standard deviations. Thus, if $\bar{X} = 100$ and $s = 10$, practically all the observations lie between $100 + 3(10)$ and $100 - 3(10)$, or 70 and 130. The estimated range is therefore 60, found by $130 - 70$.

Conversely, if we know that the range is 60, we can approximate the standard deviation by dividing the range by 6. For this illustration: $\text{range} \div 6 = 60 \div 6 = 10$, the standard deviation.

Example

A sample of the rental rates at University Park Apartments approximates a symmetrical, bell-shaped distribution. The sample mean is \$500; the standard deviation is \$20. Using the Empirical Rule, answer these questions:

1. About 68 percent of the monthly food expenditures are between what two amounts?
2. About 95 percent of the monthly food expenditures are between what two amounts?
3. Almost all of the monthly expenditures are between what two amounts?

Solution

1. About 68 percent are between \$480 and \$520, found by $\bar{X} \pm 1s = \$500 \pm 1(\$20)$.
2. About 95 percent are between \$460 and \$540, found by $\bar{X} \pm 2s = \$500 \pm 2(\$20)$.
3. Almost all (99.7 percent) are between \$440 and \$560, found by $\bar{X} \pm 3s = \$500 \pm 3(\$20)$.

Self-Review 3–9



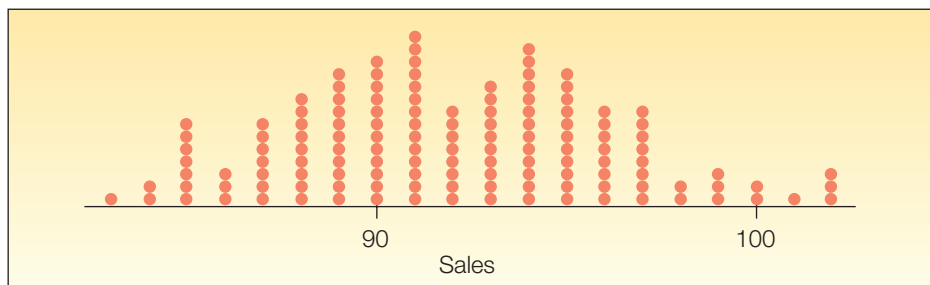
The Pitney Pipe Company is one of several domestic manufacturers of PVC pipe. The quality control department sampled 600 10-foot lengths. At a point 1 foot from the end of the pipe they measured the outside diameter. The mean was 14.0 inches and the standard deviation 0.1 inches.

- (a) If the shape of the distribution is not known, at least what percent of the observations will be between 13.85 inches and 14.15 inches?
- (b) If we assume that the distribution of diameters is symmetrical and bell-shaped, about 95 percent of the observations will be between what two values?

Exercises

connect™

53. According to Chebyshev's theorem, at least what percent of any set of observations will be within 1.8 standard deviations of the mean?
54. The mean income of a group of sample observations is \$500; the standard deviation is \$40. According to Chebyshev's theorem, at least what percent of the incomes will lie between \$400 and \$600?
55. The distribution of the weights of a sample of 1,400 cargo containers is symmetric and bell-shaped. According to the Empirical Rule, what percent of the weights will lie:
 - a. Between $\bar{X} - 2s$ and $\bar{X} + 2s$?
 - b. Between \bar{X} and $\bar{X} + 2s$? Below $\bar{X} - 2s$?
56. The following graph portrays the distribution of the number of Biggie-sized soft drinks sold at a nearby Wendy's for the last 141 days. The mean number of drinks sold per day is 91.9 and the standard deviation is 4.67.



If we use the Empirical Rule, sales will be between what two values on 68 percent of the days? Sales will be between what two values on 95 percent of the days?



Statistics in Action

Magglio Ordonez of the Detroit Tigers had the highest batting average at .363 during the 2007 season. Tony Gwynn hit .394 in the strike-shortened season of 1994, and Ted Williams hit .406 in 1941. No one has hit over .400 since 1941. The mean batting average has remained constant at about .260 for more than 100 years, but the standard deviation declined from .049 to .031. This indicates less dispersion in the batting averages today and helps explain the lack of any .400 hitters in recent times.

The Mean and Standard Deviation of Grouped Data

In most instances measures of location, such as the mean, and measures of dispersion, such as the standard deviation, are determined by using the individual values. Statistical software packages make it easy to calculate these values, even for large data sets. However, sometimes we are given only the frequency distribution and wish to estimate the mean or standard deviation. In the following discussion we show how we can estimate the mean and standard deviation from data organized into a frequency distribution. We should stress that a mean or a standard deviation from grouped data is an *estimate* of the corresponding actual values.

The Arithmetic Mean

To approximate the arithmetic mean of data organized into a frequency distribution, we begin by assuming the observations in each class are represented by the *midpoint* of the class. The mean of a sample of data organized in a frequency distribution is computed by:

$$\text{ARITHMETIC MEAN OF GROUPED DATA} \quad \bar{X} = \frac{\sum fM}{n} \quad [3-12]$$

where:

- \bar{X} is the designation for the sample mean.
- M is the midpoint of each class.
- f is the frequency in each class.
- fM is the frequency in each class times the midpoint of the class.
- $\sum fM$ is the sum of these products.
- n is the total number of frequencies.

Example

The computations for the arithmetic mean of data grouped into a frequency distribution will be shown based on the Whitner Autoplex data. Recall in Chapter 2, in Table 2-7 on page 31 we constructed a frequency distribution for the vehicle selling prices. The information is repeated below. Determine the arithmetic mean vehicle selling price.

Selling Price (\$ thousands)	Frequency
15 up to 18	8
18 up to 21	23
21 up to 24	17
24 up to 27	18
27 up to 30	8
30 up to 33	4
33 up to 36	2
Total	$\bar{80}$

Solution

The mean vehicle selling price can be estimated from data grouped into a frequency distribution. To find the estimated mean, assume the midpoint of each class is representative of the data values in that class. Recall that the midpoint of a class is halfway between the upper and the lower class limits. To find the midpoint of a particular class, we add the upper and the lower class limits and divide by 2. Hence, the midpoint of the first class is \$16.5, found by $(\$15 + \$18)/2$. We assume that the value of \$16.5 is representative of the eight values in that class. To put it another way, we assume the sum of the eight values in this class is \$132, found by $8(\$16.5)$. We continue the process of multiplying the class midpoint by the class frequency for each class and then sum these products. The results are summarized in Table 3-1.

TABLE 3-1 Price of 80 New Vehicles Sold Last Month at Whitner Autoplex Lot

Selling Price (\$ thousands)	Frequency (<i>f</i>)	Midpoint (<i>M</i>)	<i>fM</i>
15 up to 18	8	\$16.5	\$ 132.0
18 up to 21	23	19.5	448.5
21 up to 24	17	22.5	382.5
24 up to 27	18	25.5	459.0
27 up to 30	8	28.5	228.0
30 up to 33	4	31.5	126.0
33 up to 36	2	34.5	69.0
Total	80		\$1,845.0

Solving for the arithmetic mean using formula (3-12), we get:

$$\bar{X} = \frac{\sum fM}{n} = \frac{\$1,845}{80} = \$23.1 \text{ (thousands)}$$

So we conclude that the mean vehicle selling price is about \$23,100.

Standard Deviation

To calculate the standard deviation of data grouped into a frequency distribution, we need to adjust formula (3-11) slightly. We weight each of the squared differences by the number of frequencies in each class. The formula is:

STANDARD DEVIATION, GROUPED DATA

$$s = \sqrt{\frac{\sum f(M - \bar{X})^2}{n - 1}}$$

[3-13]

where:

s is the symbol for the sample standard deviation.

M is the midpoint of the class.

f is the class frequency.

\bar{n} is the number of observations in the sample.

\bar{X} is the designation for the sample mean.

Example

Refer to the frequency distribution for the Whitner Autoplex data reported in Table 3–1. Compute the standard deviation of the vehicle selling prices.

Solution

Following the same practice used earlier for computing the mean of data grouped into a frequency distribution, f is the class frequency, M the class midpoint, and n the number of observations.

Selling Price (\$ thousands)	Frequency (f)	Midpoint (M)	$(M - \bar{X})$	$(M - \bar{X})^2$	$f(M - \bar{X})^2$
15 up to 18	8	16.5	-6.6	43.56	348.48
18 up to 21	23	19.5	-3.6	12.96	298.08
21 up to 24	17	22.5	-0.6	0.36	6.12
24 up to 27	18	25.5	2.4	5.76	103.68
27 up to 30	8	28.5	5.4	29.16	233.28
30 up to 33	4	31.5	8.4	70.56	282.24
33 up to 36	2	34.5	11.4	129.96	259.92
	<u>80</u>				<u>1,531.80</u>

To find the standard deviation:

Step 1: Subtract the mean from the class midpoint. That is, find $(M - \bar{X})$. For the first class $(16.5 - 23.1 = -6.6)$, for the second class $(19.5 - 23.1 = -3.6)$, and so on.

Step 2: Square the difference between the class midpoint and the mean. For the first class it would be $(16.5 - 23.1)^2 = (-6.6)^2 = 43.56$, for the second class $(19.5 - 23.1)^2 = (-3.6)^2 = 12.96$, and so on.

Step 3: Multiply the squared difference between the class midpoint and the mean by the class frequency. For the first class the value is $8(16.5 - 23.1)^2 = 348.48$, for the second $23(19.5 - 23.1)^2 = 298.08$, and so on.

Step 4: Sum the $f(M - \bar{X})^2$. The total is 1,531.8.

To find the standard deviation we insert these values in formula (3–13).

$$s = \sqrt{\frac{\sum f(M - \bar{X})^2}{n - 1}} = \sqrt{\frac{1531.8}{80 - 1}} = 4.403.$$

The mean and the standard deviation calculated from the data grouped into a frequency distribution are usually close to the values calculated from raw data. The grouped data result in some loss of information. For the vehicle selling price problem the mean selling price reported in the Excel output on page 66 is \$23,218 and the standard deviation is \$4,354. The respective values estimated from data grouped into a frequency distribution are \$23,100 and \$4,403. The difference in the means is \$118 or about 0.51 percent. The standard deviations differ by \$49 or 1.1 percent. Based on the percentage difference, the estimates are very close to the actual values.

Self-Review 3–10

The net incomes of a sample of large importers of antiques were organized into the following table:



Net Income (\$ millions)	Number of Importers	Net Income (\$ millions)	Number of Importers
2 up to 6	1	14 up to 18	3
6 up to 10	4	18 up to 22	2
10 up to 14	10		

- (a) What is the table called?
 (b) Based on the distribution, what is the estimate of the arithmetic mean net income?
 (c) Based on the distribution, what is the estimate of the standard deviation?

connect™

Exercises

57. When we compute the mean of a frequency distribution, why do we refer to this as an *estimated* mean?
 58. Determine the mean and the standard deviation of the following frequency distribution.

Class	Frequency
0 up to 5	2
5 up to 10	7
10 up to 15	12
15 up to 20	6
20 up to 25	3

59. Determine the mean and the standard deviation of the following frequency distribution.

Class	Frequency
20 up to 30	7
30 up to 40	12
40 up to 50	21
50 up to 60	18
60 up to 70	12

60. SCCoast, an Internet provider in the Southeast, developed the following frequency distribution on the age of Internet users. Find the mean and the standard deviation.

Age (years)	Frequency
10 up to 20	3
20 up to 30	7
30 up to 40	18
40 up to 50	20
50 up to 60	12

61. The IRS was interested in the number of individual tax forms prepared by small accounting firms. The IRS randomly sampled 50 public accounting firms with 10 or fewer employees in the Dallas–Fort Worth area. The following frequency table reports the results of the study. Estimate the mean and the standard deviation.

Number of Clients	Frequency
20 up to 30	1
30 up to 40	15
40 up to 50	22
50 up to 60	8
60 up to 70	4

62. Advertising expenses are a significant component of the cost of goods sold. Listed below is a frequency distribution showing the advertising expenditures for 60 manufacturing companies located in the Southwest. Estimate the mean and the standard deviation of advertising expenses.

Advertising Expenditure (\$ millions)	Number of Companies
25 up to 35	5
35 up to 45	10
45 up to 55	21
55 up to 65	16
65 up to 75	8
Total	60

Ethics and Reporting Results

In Chapter 1, we discussed the ethical and unbiased reporting of statistical results. While you are learning about how to organize, summarize, and interpret data using statistics, it is also important to understand statistics so that you can be an intelligent consumer of information.

In this chapter, we learned how to compute numerical descriptive statistics. Specifically, we showed how to compute and interpret measures of location for a data set: the mean, median, and mode. We also discussed the advantages and disadvantages for each statistic. For example, if a real estate developer tells a client that the average home in a particular subdivision sold for \$150,000, we assume that \$150,000 is a representative selling price for all the homes. But suppose that the client also asks what the median sales price is, and the median is \$60,000. Why was the developer only reporting the mean price? This information is extremely important to a person's decision making when buying a home. Knowing the advantages and disadvantages of the mean, median, and mode is important as we report statistics and as we use statistical information to make decisions.

We also learned how to compute measures of dispersion: range, mean deviation, and standard deviation. Each of these statistics also has advantages and disadvantages. Remember that the range provides information about the overall spread of a distribution. However, it does not provide any information about how the data is clustered or concentrated around the center of the distribution.

As we learn more about statistics, we need to remember that when we use statistics we must maintain an independent and principled point of view. Any statistical report requires objective and honest communication of the results.

Chapter Summary

- I. A measure of location is a value used to describe the center of a set of data.
 - A. The arithmetic mean is the most widely reported measure of location.
 1. It is calculated by adding the values of the observations and dividing by the total number of observations.
 - a. The formula for a population mean of ungrouped or raw data is

$$\mu = \frac{\sum X}{N}$$

[3-1]

- b. The formula for the mean of a sample is

$$\bar{X} = \frac{\sum X}{n} \quad [3-2]$$

- c. The formula for the sample mean of data in a frequency distribution is

$$\bar{X} = \frac{\sum fM}{n} \quad [3-12]$$

2. The major characteristics of the arithmetic mean are:
- At least the interval scale of measurement is required.
 - All the data values are used in the calculation.
 - A set of data has only one mean. That is, it is unique.
 - The sum of the deviations from the mean equals 0.
- B. The weighted mean is found by multiplying each observation by its corresponding weight.
1. The formula for determining the weighted mean is

$$\bar{X}_w = \frac{w_1X_1 + w_2X_2 + w_3X_3 + \cdots + w_nX_n}{w_1 + w_2 + w_3 + \cdots + w_n} \quad [3-3]$$

2. It is a special case of the arithmetic mean.
- C. The median is the value in the middle of a set of ordered data.
- To find the median, sort the observations from smallest to largest and identify the middle value.
 - The major characteristics of the median are:
 - At least the ordinal scale of measurement is required.
 - It is not influenced by extreme values.
 - Fifty percent of the observations are larger than the median.
 - It is unique to a set of data.
- D. The mode is the value that occurs most often in a set of data.
- The mode can be found for nominal-level data.
 - A set of data can have more than one mode.
- E. The geometric mean is the n th root of the product of n positive values.
1. The formula for the geometric mean is

$$GM = \sqrt[n]{(X_1)(X_2)(X_3) \cdots (X_n)} \quad [3-4]$$

2. The geometric mean is also used to find the rate of change from one period to another.

$$GM = \sqrt[n]{\frac{\text{Value at end of period}}{\text{Value at beginning of period}}} - 1 \quad [3-5]$$

3. The geometric mean is always equal to or less than the arithmetic mean.
- II. The dispersion is the variation or spread in a set of data.
- A. The range is the difference between the largest and the smallest value in a set of data.
1. The formula for the range is

$$\text{Range} = \text{Largest value} - \text{Smallest value} \quad [3-6]$$

2. The major characteristics of the range are:
- Only two values are used in its calculation.
 - It is influenced by extreme values.
 - It is easy to compute and to understand.
- B. The mean absolute deviation is the sum of the absolute values of the deviations from the mean divided by the number of observations.
1. The formula for computing the mean absolute deviation is

$$MD = \frac{\sum |X - \bar{X}|}{n} \quad [3-7]$$

2. The major characteristics of the mean absolute deviation are:
- It is not unduly influenced by large or small values.
 - All observations are used in the calculation.
 - The absolute values are somewhat difficult to work with.

- C. The variance is the mean of the squared deviations from the arithmetic mean.
1. The formula for the population variance is

$$\sigma^2 = \frac{\sum(X - \mu)^2}{N} \quad [3-8]$$

2. The formula for the sample variance is

$$s^2 = \frac{\sum(X - \bar{X})^2}{n - 1} \quad [3-10]$$

3. The major characteristics of the variance are:
 - a. All observations are used in the calculation.
 - b. It is not unduly influenced by extreme observations.
 - c. The units are somewhat difficult to work with; they are the original units squared.

- D. The standard deviation is the square root of the variance.

1. The major characteristics of the standard deviation are:
 - a. It is in the same units as the original data.
 - b. It is the square root of the average squared distance from the mean.
 - c. It cannot be negative.
 - d. It is the most widely reported measure of dispersion.
2. The formula for the sample standard deviation is

$$s = \sqrt{\frac{\sum(X - \bar{X})^2}{n - 1}} \quad [3-11]$$

3. The formula for the standard deviation of grouped data is

$$s = \sqrt{\frac{\sum f(M - \bar{X})^2}{n - 1}} \quad [3-13]$$

- III. We interpret the standard deviation using two measures.

- A. Chebyshev's theorem states that regardless of the shape of the distribution, at least $1 - 1/k^2$ of the observations will be within k standard deviations of the mean, where k is greater than 1.
- B. The Empirical Rule states that for a bell-shaped distribution about 68 percent of the values will be within one standard deviation of the mean, 95 percent within two, and virtually all within three.

Pronunciation Key

SYMBOL	MEANING	PRONUNCIATION
μ	Population mean	<i>mu</i>
Σ	Operation of adding	<i>sigma</i>
ΣX	Adding a group of values	<i>sigma X</i>
\bar{X}	Sample mean	<i>X bar</i>
\bar{X}_w	Weighted mean	<i>X bar sub w</i>
<i>GM</i>	Geometric mean	<i>G M</i>
ΣfM	Adding the product of the frequencies and the class midpoints	<i>sigma f M</i>
σ^2	Population variance	<i>sigma squared</i>
σ	Population standard deviation	<i>sigma</i>

Chapter Exercises

connect™

63. The accounting firm of Crawford and Associates has five senior partners. Yesterday the senior partners saw six, four, three, seven, and five clients, respectively.
- a. Compute the mean number and median number of clients seen by the partners.
 - b. Is the mean a sample mean or a population mean?
 - c. Verify that $\sum(X - \mu) = 0$.

64. Owens Orchards sells apples in a large bag by weight. A sample of seven bags contained the following numbers of apples: 23, 19, 26, 17, 21, 24, 22.
 a. Compute the mean number and median number of apples in a bag.
 b. Verify that $\Sigma(X - \bar{X}) = 0$.
65. A sample of households that subscribe to United Bell Phone Company for land line phone service revealed the following number of calls received per household last week. Determine the mean and the median number of calls received.

52	43	30	38	30	42	12	46	39	37
34	46	32	18	41	5				

66. The Citizens Banking Company is studying the number of times the ATM located in a Loblaws Supermarket at the foot of Market Street is used per day. Following are the number of times the machine was used daily over each of the last 30 days. Determine the mean number of times the machine was used per day.

83	64	84	76	84	54	75	59	70	61
63	80	84	73	68	52	65	90	52	77
95	36	78	61	59	84	95	47	87	60

67. The Canadian government wants to know the relative age of its workforce. As the baby boom generation becomes older, the government is concerned about the availability of younger qualified workers. To become more informed, the government surveyed many industries regarding employee ages. The mean and median age for two industries, communication and retail trade, for six different job types are listed in the following table.

	Communication and Other Utilities		Retail Trade and Consumer Services	
	Mean	Median	Mean	Median
Managers	42.6	43	38.6	38
Professionals	40.8	40	40.0	39
Technical/Trades	41.4	42	37.1	37
Marketing/Sales	NA	NA	33.7	31
Clerical/Administrative	40.8	41	38.0	38
Production Workers	37.2	40	32.0	24

Comment on the distribution of age. Which industry appears to have older workers? Younger workers? In each industry, which job types show the greatest difference between mean and median age?

68. Trudy Green works for the True-Green Lawn Company. Her job is to solicit lawn-care business via the telephone. Listed below is the number of appointments she made in each of the last 25 hours of calling. What is the arithmetic mean number of appointments she made per hour? What is the median number of appointments per hour? Write a brief report summarizing the findings.

9	5	2	6	5	6	4	4	7	2	3	6	3
4	4	7	8	4	4	5	5	4	8	3	3	

69. The Split-A-Rail Fence Company sells three types of fence to homeowners in suburban Seattle, Washington. Grade A costs \$5.00 per running foot to install, Grade B costs \$6.50 per running foot, and Grade C, the premium quality, costs \$8.00 per running foot. Yesterday, Split-A-Rail installed 270 feet of Grade A, 300 feet of Grade B, and 100 feet of Grade C. What was the mean cost per foot of fence installed?
70. Rolland Poust is a sophomore in the College of Business at Scandia Tech. Last semester he took courses in statistics and accounting, 3 hours each, and earned an A in both. He earned a B in a five-hour history course and a B in a two-hour history of jazz course. In addition, he took a one-hour course dealing with the rules of basketball so he could get his license to officiate high school basketball games. He got an A in this course. What was his GPA for the semester? Assume that he receives 4 points for an A, 3 for a B, and so on. What measure of location did you just calculate?

71. The table below shows the percent of the labor force that is unemployed and the size of the labor force for three counties in Northwest Ohio. Jon Elsas is the Regional Director of Economic Development. He must present a report to several companies that are considering locating in Northwest Ohio. What would be an appropriate unemployment rate to show for the entire region?

County	Percent Unemployed	Size of Workforce
Wood	4.5	15,300
Ottawa	3.0	10,400
Lucas	10.2	150,600

72. The American Automobile Association checks prices of gasoline before many holiday weekends. Listed below are the self-service prices for a sample of 15 retail outlets during the last Labor Day weekend in the Detroit, Michigan area.

3.44	3.42	3.35	3.39	3.49	3.49	3.41	3.46
3.41	3.49	3.45	3.48	3.39	3.46	3.44	

- a. What is the arithmetic mean selling price?
 b. What is the median selling price?
 c. What is the modal selling price?
73. The metropolitan area of Los Angeles–Long Beach, California, is the area expected to show the largest increase in the number of jobs between 1989 and 2010. The number of jobs is expected to increase from 5,164,900 to 6,286,800. What is the geometric mean expected yearly rate of increase?
74. A recent article suggested that, if you earn \$25,000 a year today and the inflation rate continues at 3 percent per year, you'll need to make \$33,598 in 10 years to have the same buying power. You would need to make \$44,771 if the inflation rate jumped to 6 percent. Confirm that these statements are accurate by finding the geometric mean rate of increase.
75. The ages of a sample of Canadian tourists flying from Toronto to Hong Kong were: 32, 21, 60, 47, 54, 17, 72, 55, 33, and 41.
 a. Compute the range.
 b. Compute the mean deviation.
 c. Compute the standard deviation.
76. The weights (in pounds) of a sample of five boxes being sent by UPS are: 12, 6, 7, 3, and 10.
 a. Compute the range.
 b. Compute the mean deviation.
 c. Compute the standard deviation.
77. A southern state has seven state universities in its system. The numbers of volumes (in thousands) held in its libraries are 83, 510, 33, 256, 401, 47, and 23.
 a. Is this a sample or a population?
 b. Compute the standard deviation.
78. Health issues are a concern of managers, especially as they evaluate the cost of medical insurance. A recent survey of 150 executives at Elvers Industries, a large insurance and financial firm located in the Southwest, reported the number of pounds by which the executives were overweight. Compute the mean and the standard deviation.

Pounds Overweight	Frequency
0 up to 6	14
6 up to 12	42
12 up to 18	58
18 up to 24	28
24 up to 30	8

79. The Apollo space program lasted from 1967 until 1972 and included 13 missions. The missions lasted from as little as 7 hours to as long as 301 hours. The duration of each flight is listed below.

9	195	241	301	216	260	7	244	192	147
10	295	142							

- a. Explain why the flight times are a population.
 - b. Find the mean and median of the flight times.
 - c. Find the range and the standard deviation of the flight times.
80. Creek Ratz is a very popular restaurant located along the coast of northern Florida. They serve a variety of steak and seafood dinners. During the summer beach season, they do not take reservations or accept “call ahead” seating. Management of the restaurant is concerned with the time a patron must wait before being seated for dinner. Listed below is the wait time, in minutes, for the 25 tables seated last Saturday night.

28	39	23	67	37	28	56	40	28	50
51	45	44	65	61	27	24	61	34	44
64	25	24	27	29					

- a. Explain why the times are a population.
 - b. Find the mean and median of the times.
 - c. Find the range and the standard deviation of the times.
81. A sample of 25 undergraduates reported the following dollar amounts of entertainment expenses last year:

684	710	688	711	722	698	723	743	738	722	696	721	685
763	681	731	736	771	693	701	737	717	752	710	697	

- a. Find the mean, median and mode of this information.
 - b. What are the range and standard deviation?
 - c. Use the Empirical Rule to establish an interval which includes about 95 percent of the observations.
82. The EPA recently released the following air quality indexes for selected metropolitan areas.

Index	Metropolitan Area	Index	Metropolitan Area
41	Allentown–Bethlehem–Easton, PA	29	Monroe, LA
45	Athens, GA	48	New York, NY
44	Bergen–Passaic, NJ	34	Oakland, CA
39	Buffalo–Niagara Falls, NY	31	Pittsfield, MA
32	Cedar Rapids, IA	36	Reading, PA
42	Cleveland–Lorain–Elyria, OH	31	Rochester, NY
37	Florence, AL	38	San Antonio, TX
41	Fort Worth–Arlington, TX	42	Savannah, GA
37	Goldsboro, NC	35	Scranton–Wilkes Barre–Hazleton, PA
38	Huntsville, AL	36	Vineland–Millville–Bridgeton, NJ
38	Jacksonville, FL	54	Visalia–Tulare–Porterville, CA
33	Johnstown, PA	47	Wheeling, WV–OH
20	Mayaguez, PR		

- a. Find the mean, median, and mode of this data set.
 - b. What are the range and standard deviation of the readings?
 - c. Use the Empirical Rule to establish an interval which includes about 95 percent of the observations.
83. U.S. housing prices reached their peak in July of 2006. The following data shows the ratio of the prices reported in a recent month to the July 2006 figure. These are to be used to represent typical amounts of decline from that peak for 20 cities.

Phoenix	Los Angeles	San Diego	San Francisco	Denver	Washington	Miami	Tampa	Atlanta	Chicago
0.760	0.784	0.764	0.802	0.909	0.828	0.786	0.792	0.933	0.915
Boston	Detroit	Minneapolis	Charlotte	Las Vegas	New York	Cleveland	Portland	Dallas	Seattle
0.902	0.792	0.855	1.025	0.756	0.922	0.865	0.981	0.939	0.999

- a. Find the mean, median, and mode of these statistics.
 - b. What are the range and standard deviation of the values? (Assume this to be sample information.)
 - c. Use the Empirical Rule to determine an interval which contains approximately 95 percent of the observations.
84. The Kentucky Derby is held the first Saturday in May at Churchill Downs in Louisville, KY. The racing track is one and one-quarter miles. The following table shows the winners since 1990, their margin of victory, the winning time, and the payoff on a \$2 bet.

Year	Winner	Margin (lengths)	Time (minutes)	Payoffs for \$2 Bets
1990	Unbridled	3.5	2.03333	10.80
1991	Strike the Gold	1.75	2.05000	4.80
1992	Lil E. Tee	1	2.05000	16.80
1993	Sea Hero	2.5	2.04000	12.90
1994	Go For Gin	2	2.06000	9.10
1995	Thunder Gulch	2.25	2.02000	24.50
1996	Grindstone	Nose	2.01667	5.90
1997	Silver Charm	Head	2.04000	4.00
1998	Real Quiet	0.5	2.03667	8.40
1999	Charismatic	Neck	2.05333	31.30
2000	Fusaichi Pegasus	1.5	2.02000	2.30
2001	Monarchos	4.75	1.99667	10.50
2002	War Emblem	4	2.01667	20.50
2003	Funny Cide	1.75	2.01667	12.80
2004	Smarty Jones	2.75	2.06667	4.10
2005	Giacomo	0.5	2.04583	50.30
2006	Barbaro	6.5	2.03933	6.10
2007	Street Sense	2.25	2.03617	4.90
2008	Big Brown	4.75	2.03033	6.80

- a. Obtain the mean and median for the variables winning time and payoff on a \$2 bet.
 - b. Determine the range and standard deviation of the variables time and payoff.
 - c. Refer to the variable winning margin. What is the level of measurement? What measure of location would be most appropriate?
85. The manager of the local Wal-Mart Super Store is studying the number of items purchased by customers in the evening hours. Listed below is the number of items for a sample of 30 customers.

15	8	6	9	9	4	18	10	10	12
12	4	7	8	12	10	10	11	9	13
5	6	11	14	5	6	6	5	13	5

- a. Find the mean and the median of the number of items.
 - b. Find the range and the standard deviation of the number of items.
 - c. Organize the number of items into a frequency distribution. You may want to review the guidelines in Chapter 2 for establishing the class interval and the number of classes.
 - d. Find the mean and the standard deviation of the data organized into a frequency distribution. Compare these values with those computed in part (a). Why are they different?
86. The following frequency distribution reports the electricity cost for a sample of 50 two-bedroom apartments in Albuquerque, New Mexico during the month of May last year.

Electricity Cost	Frequency
\$ 80 up to \$100	3
100 up to 120	8
120 up to 140	12
140 up to 160	16
160 up to 180	7
180 up to 200	4
Total	50

- a. Estimate the mean cost.
 - b. Estimate the standard deviation.
 - c. Use the Empirical Rule to estimate the proportion of costs within two standard deviations of the mean. What are these limits?
87. Bidwell Electronics, Inc., recently surveyed a sample of employees to determine how far they lived from corporate headquarters. The results are shown below. Compute the mean and the standard deviation.

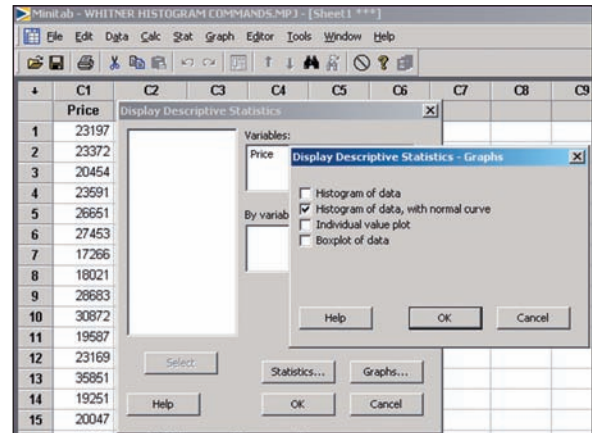
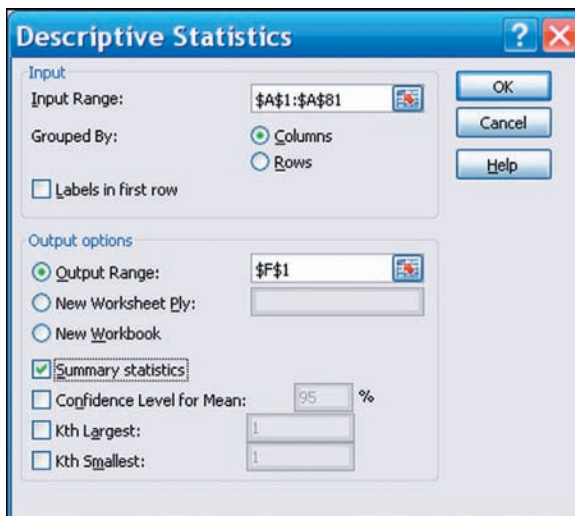
Distance (miles)	Frequency	<i>M</i>
0 up to 5	4	2.5
5 up to 10	15	7.5
10 up to 15	27	12.5
15 up to 20	18	17.5
20 up to 25	6	22.5

Data Set Exercises

88. Refer to the Real Estate data, which reports information on homes sold in the Denver, Colorado, area last year.
- a. Select the variable selling price.
 1. Find the mean, median, and the standard deviation.
 2. Write a brief summary of the distribution of selling prices.
 - b. Select the variable referring to the area of the home in square feet.
 1. Find the mean, median, and the standard deviation.
 2. Write a brief summary of the distribution of the area of homes.
89. Refer to the Baseball 2007 data, which reports information on the 30 major league teams for the 2007 baseball season.
- a. Select the variable team salary and find the mean, median, and the standard deviation.
 - b. Select the variable that refers to the age the stadium was built. (Hint: Subtract the year in which the stadium was built from the current year to find the stadium age and work with that variable.) Find the mean, median, and the standard deviation.
 - c. Select the variable that refers to the seating capacity of the stadium. Find the mean, median, and the standard deviation.
90. Refer to the CIA data, which reports demographic and economic information on 46 countries.
- a. Select the variable life expectancy.
 1. Find the mean, median, and the standard deviation.
 2. Write a brief summary of the distribution of life expectancy.
 - b. Select the variable GDP/cap.
 1. Find the mean, median, and the standard deviation.
 2. Write a brief summary of the distribution GDP/cap.

Software Commands

- The Excel Commands for the descriptive statistics on page 66 are:
 - From the CD retrieve the Whitner data file, which is called **Whitner**.
 - From the menu bar select **Data** and then **Data Analysis**. Select **Descriptive Statistics** and then click **OK**.
 - For the **Input Range**, type **A1:A81**, indicate that the data are grouped by column and that the labels are in the first row. Click on **Output Range**, indicate that the output should go in **H1** (or any place you wish), click on **Summary statistics**, then click **OK**.
 - After you get your results, double-check the count in the output to be sure it contains the correct number of items.
- The MINITAB commands for the descriptive summary on page 81 are:
 - From the CD retrieve the Whitner data, which is called **Whitner**.
 - Select **Stat, Basic Statistics**, and then **Display Descriptive Statistics**. In the dialog box select **Price** as the variable and then click on **Graphs** in the lower right-hand corner. Within the new dialog box select **Histogram of data, with normal curve** and click **OK**. Click **OK** in the next dialog box.



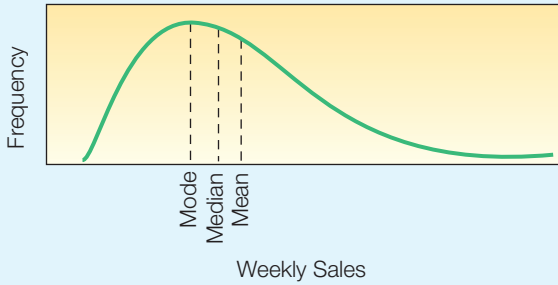
Chapter 3 Answers to Self-Review



- 3-1
- $\bar{X} = \frac{\sum X}{n}$
 - $\bar{X} = \frac{\$267,100}{4} = \$66,775$
 - Statistic, because it is a sample value.
 - \$66,775. The sample mean is our best estimate of the population mean.
 - $\mu = \frac{\sum X}{N}$
 - $\mu = \frac{498}{6} = 83$
 - Parameter, because it was computed using all the population values.
- 3-2
- \$237, found by:

$$\frac{(95 \times \$400) + (126 \times \$200) + (79 \times \$100)}{95 + 126 + 79} = \$237.00$$
 - The profit per suit is \$12, found by \$237 – \$200 cost – \$25 commission. The total profit for the 300 suits is \$3,600, found by $300 \times \$12$.
- 3-3
- \$878
 - 3, 3
 - 7, found by $(6 + 8)/2 = 7$
 - 3, 3
 - 0

3-4 a.



b. Positively skewed, because the mean is the largest average and the mode is the smallest.

- 3-5 1. a. About 9.9 percent, found by $\sqrt[4]{1.458602236}$, then $1.099 - 1.00 = .099$
 b. About 10.095 percent
 c. Greater than, because $10.095 > 9.9$

2. 8.63 percent, found by $\sqrt[20]{\frac{120,520}{23,000}} - 1 = 1.0863 - 1$

- 3-6 a. 22 thousands of pounds, found by $112 - 90$
 b. $\bar{X} = \frac{824}{8} = 103$ thousands of pounds

c.

X	$ X - \bar{X} $	Absolute Deviation
95	8	8
103	0	0
105	2	2
110	7	7
104	1	1
105	2	2
112	9	9
90	13	13
		Total 42

$$MD = \frac{42}{8} = 5.25 \text{ thousands of pounds}$$

- 3-7 a. $\mu = \frac{\$16,900}{5} = \$3,380$
 b. $\sigma^2 = \frac{(3536 - 3380)^2 + \dots + (3622 - 3380)^2}{5}$

$$= \frac{(156)^2 + (-207)^2 + (68)^2 + (-259)^2 + (242)^2}{5}$$

$$= \frac{197,454}{5} = 39,490.8$$

 c. $\sigma = \sqrt{39,490.8} = 198.72$

d. There is more variation in the Pittsburgh office because the standard deviation is larger. The mean is also larger in the Pittsburgh office.

3-8 2.33, found by:

$$\bar{X} = \frac{\Sigma X}{n} = \frac{28}{7} = 4$$

X	$X - \bar{X}$	$(X - \bar{X})^2$
4	0	0
2	-2	4
5	1	1
4	0	0
5	1	1
2	-2	4
6	2	4
28	0	14

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

$$= \frac{14}{7 - 1}$$

$$= 2.33$$

$$s = \sqrt{2.33} = 1.53$$

- 3-9 a. $k = \frac{14.15 - 14.00}{.10} = 1.5$
 $k = \frac{13.85 - 14.0}{.10} = -1.5$
 $1 - \frac{1}{(1.5)^2} = 1 - .44 = .56$

b. 13.8 and 14.2

3-10 a. Frequency distribution.

b.

f	M	fM	$(M - \bar{X})$	$f(M - \bar{X})^2$
1	4	4	-8.2	67.24
4	8	32	-4.2	70.56
10	12	120	-0.2	0.40
3	16	48	3.8	43.32
2	20	40	7.8	121.68
20		244		303.20

$$\bar{X} = \frac{\Sigma fM}{M} = \frac{\$244}{20} = \$12.20$$

$$c. s = \sqrt{\frac{303.20}{20 - 1}} = \$3.99$$