3
Descriptive Statistics

Learning Objectives

The focus of Chapter 3 is the use of statistical techniques to describe data, thereby enabling you to:

1. Distinguish between measures of central tendency, measures of variability, and measures of shape.
2. Understand the meanings of mean, median, mode, quartile, and range.
3. Compute mean, median, mode, quartile, range, variance, standard deviation, and mean absolute deviation.
4. Differentiate between sample and population variance and standard deviation.
5. Understand the meaning of standard deviation as it is applied by using the empirical rule.
6. Understand box and whisker plots, skewness, and kurtosis.

Chapter 2 described graphical techniques for organizing and presenting data. While these graphs allow the researcher to make some general observations about the shape and spread of the data, a fuller understanding of the data can be attained by summarizing the data numerically using statistics. This chapter presents such statistical measures, including measures of central tendency, measures of variability, and measures of shape.

3.1 Measures of Central Tendency

One type of measure that is used to describe a set of data is the **measure of central tendency.** Measures of central tendency *yield information about the center, or middle part, of a group of numbers.* Displayed in Table 3.1 are the offer price for the 20 largest U.S. initial public offerings in a recent year according to the Securities Data Co. For these data, measures of central tendency can yield such information as the average offer price, the middle offer price, and the most frequently occurring offer price. Measures of central tendency do not focus on the span of the data set or how far values are from the middle numbers. The measures of central tendency presented here for ungrouped data are the mode, the median, the mean, and quartiles.

Measure of central tendency
One type of measure that is used to yield information about the center of a group of numbers.

Mode

The **mode** is *the most frequently occurring value* in a set of data. For the data in Table 3.1 the mode is $19.00 because the offer price that recurred the most times (4) was $19.00. Organizing the data into an *ordered array* (an ordering of the numbers from smallest to largest) helps to locate the mode. The following is an ordered array of the values from Table 3.1.

Mode
The most frequently occurring value in a set of data.

7.00	11.00	14.25	15.00	15.00	15.50	19.00	19.00	19.00	19.00
21.00	22.00	23.00	24.00	25.00	27.00	27.00	28.00	34.22	43.25

This grouping makes it easier to see that 19.00 is the most frequently occurring number.

If there is a tie for the most frequently occurring value, there are two modes. In that case the data are said to be **bimodal.** If a set of data is not exactly bimodal but contains two values that are more dominant than others, some researchers take the liberty of referring to the data set as bimodal even though there is not an exact tie for the mode. Data sets with more than two modes are referred to as **multimodal.**

Bimodal
Data sets that have two modes.

Multimodal
Data sets that contain more than two modes.

In the world of business, the concept of mode is often used in determining sizes. For example, shoe manufacturers might produce inexpensive shoes in three widths only: small, medium, and large. Each width size represents a modal width of feet. By reducing the number of sizes to a few modal sizes, companies can reduce total product costs by limiting machine setup costs. Similarly, the garment industry produces shirts, dresses, suits, and many other clothing products in modal sizes. For example, all size M shirts in a given lot are produced in the same size. This size is some modal size for medium-size men.

The mode is an appropriate measure of central tendency for nominal level data. The mode can be used to determine which category occurs most frequently.

TABLE 3.1
Offer Prices for the Twenty Largest U.S. Initial Public Offerings in a Recent Year ($)

14.25	19.00	11.00	28.00
24.00	23.00	43.25	19.00
27.00	25.00	15.00	7.00
34.22	15.50	15.00	22.00
19.00	19.00	27.00	21.00

Median

The **median** is *the middle value in an ordered array of numbers.* If there is an odd number of terms in the array, the median is the middle number. If there is an even number of terms, the median is the average of the two middle numbers. The following steps are used to determine the median.

Median
The middle value in an ordered array of numbers.

STEP **1.** Arrange the observations in an ordered data array.
STEP **2.** If there is an odd number of terms, find the middle term of the ordered array. It is the median.
STEP **3.** If there is an even number of terms, find the average of the middle two terms. This average is the median.

Suppose a business analyst wants to determine the median for the following numbers.

15 11 14 3 21 17 22 16 19 16 5 7 19 8 9 20 4

He or she arranges the numbers in an ordered array.

3 4 5 7 8 9 11 14 15 16 16 17 19 19 20 21 22

There are 17 terms (an odd number of terms), so the median is the middle number, or 15.

If the number 22 is eliminated from the list, there are only 16 terms.

3 4 5 7 8 9 11 14 15 16 16 17 19 19 20 21

Now there is an even number of terms, and the business analyst determines the median by averaging the two middle values, 14 and 15. The resulting median value is 14.5.

Another way to locate the median is by finding the $(n + 1)/2$ term in an ordered array. For example, if a data set contains 77 terms, the median is the 39th term. That is,

$$\frac{n+1}{2} = \frac{77+1}{2} = \frac{78}{2} = 39\text{th term.}$$

This formula is helpful when a large number of terms must be manipulated.

Consider the offer price data in Table 3.1. Because there are 20 values and therefore $n = 20$, the median for these data is located at the $(20 + 1)/2$ term, or the 10.5th term. This indicates that the median is located halfway between the 10th and 11th term or the average of 19.00 and 21.00. Thus, the median offer price for the largest twenty U.S. initial public offerings is $20.00.

The median is unaffected by the magnitude of extreme values. This characteristic is an advantage, because large and small values do not inordinately influence the median. For this reason, the median is often the best measure of location to use in the analysis of variables such as house costs, income, and age. Suppose, for example, that a real estate broker wants to determine the median selling price of 10 houses listed at the following prices.

$67,000
91,000
95,000
$105,000
116,000
122,000
$148,000
167,000
189,000
$5,250,000

The median is the average of the two middle terms, $116,000 and $122,000, or $119,000. This price is a reasonable representation of the prices of the 10 houses. Note that the house priced at $5,250,000 did not enter into the analysis other than to count as one of the 10 houses. If the price of the tenth house were $200,000, the results would be the same. However, if all the house prices were averaged, the resulting average price of the original 10 houses would be $635,000, higher than nine of the 10 individual prices.

A disadvantage of the median is that not all the information from the numbers is used. That is, information about the specific asking price of the most expensive house does not really enter into the computation of the median. The level of data measurement must be at least ordinal for a median to be meaningful.

Mean

Arithmetic mean
The average of a group of numbers.

The **arithmetic mean** is synonymous with the *average of a group of numbers* and is computed by summing all numbers and dividing by the number of numbers. Because the arithmetic mean is so widely used, most statisticians refer to it simply as the *mean.*

The population mean is represented by the Greek letter mu (μ). The sample mean is represented by $\overline{X}$. The formulas for computing the population mean and the sample mean are given in the boxes that follow.

POPULATION MEAN

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

SAMPLE MEAN

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

The capital Greek letter sigma (Σ) is commonly used in mathematics to represent a summation of all the numbers in a grouping.* Also, N is the number of terms in the population, and n is the number of terms in the sample. The algorithm for computing a mean is to sum all the numbers in the population or sample and divide by the number of terms.

A more formal definition of the mean is

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}.$$

However, for the purposes of this text,

$$\Sigma X \text{ denotes} \sum_{i=1}^{N} X_i.$$

It is inappropriate to use the mean to analyze data that are not at least interval level in measurement.

Suppose a company has five departments with 24, 13, 19, 26, and 11 workers each. The *population mean* number of workers in each department is 18.6 workers. The computations follow.

$$\begin{array}{r} 24 \\ 13 \\ 19 \\ 26 \\ \underline{11} \\ \Sigma X = 93 \end{array}$$

*The mathematics of summations is not discussed here. A more detailed explanation is given on the CD-ROM.

and

$$\mu = \frac{\Sigma X}{N} = \frac{93}{5} = 18.6.$$

The calculation of a sample mean uses the same algorithm as for a population mean and will produce the same answer if computed on the same data. However, it is inappropriate to compute a sample mean for a population or a population mean for a sample. Since both populations and samples are important in statistics, a separate symbol is necessary for the population mean and for the sample mean.

DEMONSTRATION PROBLEM 3.1

The number of U.S. cars in service by top car rental companies in a recent year according to *Auto Rental News* follows.

COMPANY	NUMBER OF CARS IN SERVICE
Enterprise	355,000
Hertz	250,000
Avis	200,000
National	145,000
Alamo	130,000
Budget	125,000
Dollar	62,000
FRCS (Ford)	53,150
Thrifty	34,000
Republic Replacement	32,000
DRAC (Chrysler)	27,000
U-Save	12,000
Rent-a-Wreck	12,000
Payless	12,000
Advantage	9,000

Compute the mode, the median, and the mean.

SOLUTION

Mode: 12,000

Median: There are 15 different companies in this group, so $n = 15$. The median is located at the $(15 + 1)/2 = 8$th position. Since the data are already ordered, the 8th term is 53,150, which is the median.

Mean: The total number of cars in service is $1{,}458{,}150 = \Sigma X$

$$\mu = \frac{\Sigma X}{n} = \frac{1{,}458{,}150}{15} = 97{,}210$$

The mean is affected by each and every value, which is an advantage. The mean uses all the data and each data item influences the mean. It is also a disadvantage, because extremely large or small values can cause the mean to be pulled toward the extreme value. Recall the preceding discussion of the 10 house prices. If the mean is computed on the 10 houses, the mean price is higher than the prices of nine of the houses because the \$5,250,000 house is included in the calculation. The total price of the 10 houses is \$6,350,000, and the mean price is

$$\bar{X} = \frac{\Sigma X}{n} = \frac{\$6{,}350{,}000}{10} = \$635{,}000.$$

The mean is the most commonly used measure of location because it uses each data item in its computation, it is a familiar measure, and it has mathematical properties that make it attractive to use in inferential statistics analysis.

Quartiles

Quartiles
Measures of central tendency that divide a group of data into four subgroups or parts.

Quartiles are *measures of central tendency that divide a group of data into four subgroups or parts.* There are three quartiles, denoted as Q_1, Q_2, and Q_3. The first quartile, Q_1, separates the first, or lowest, one-fourth of the data from the upper three-fourths. The second quartile, Q_2, separates the second quarter of the data from the third quarter and equals the median of the data. The third quartile, Q_3, divides the first three-quarters of the data from the last quarter. These three quartiles are shown in Figure 3.1.

Shown next is a summary of the steps used in determining the location of a quartile.

STEPS IN DETERMINING THE LOCATION OF A QUARTILE

1. Organize the numbers into an ascending-order array.
2. Calculate the quartile location (i) by:

$$i = \frac{Q}{4}(n)$$

where:
Q = the quartile of interest,
i = quartile location, and
n = number in the data set.

3. Determine the location by either (a) or (b).
   - **a.** If i *is* a whole number, quartile Q is the average of the value at the ith location and the value at the $(i + 1)$st location.
   - **b.** If i *is not* a whole number, quartile Q value is located at the whole number part of $i + 1$.

Suppose we want to determine the values of Q_1, Q_2, and Q_3 for the following numbers.

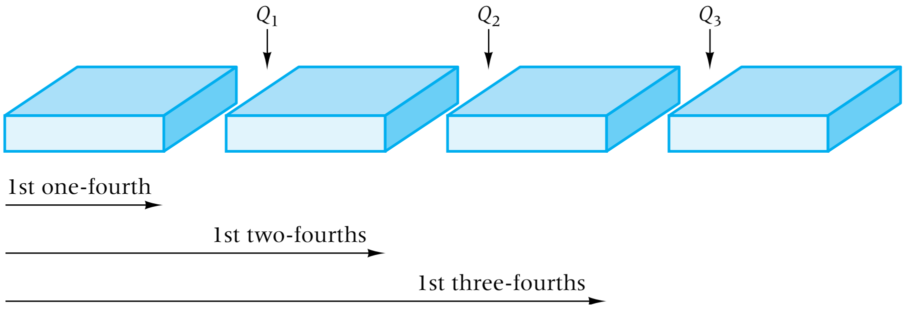
106 109 114 116 121 122 125 129

The value of Q_1 is found by

$$\text{For } n = 8, i = \frac{1}{4}(8) = 2$$

Figure 3.1

Quartiles



Because i is a whole number, Q_1 is found as the average of the second and third numbers.

$$Q_1 = \frac{(109 + 114)}{2} = 111.5$$

The value of Q_1 is 111.5. Notice that one-fourth, or two, of the values (106 and 109) are less than 111.5.

The value of Q_2 is equal to the median. As there is an even number of terms, the median is the average of the two middle terms.

$$Q_2 = \text{median} = \frac{(116 + 121)}{2} = 118.5$$

Notice that exactly half of the terms are less than Q_2 and half are greater than Q_2.

The value of Q_3 is determined as follows.

$$i = \frac{3}{4}(8) = 6$$

Because i is a whole number, Q_3 is the average of the sixth and the seventh numbers.

$$Q_3 = \frac{(122 + 125)}{2} = 123.5$$

The value of Q_3 is 123.5. Notice that three-fourths, or six, of the values are less than 123.5 and two of the values are greater than 123.5.

DEMONSTRATION PROBLEM 3.2

The following shows revenues for the world's top 20 advertising organizations according to *Advertising Age,* Crain Communications, Inc. Determine the first, the second, and the third quartiles for these data.

AD ORGANIZATION	HEADQUARTERS	WORLDWIDE GROSS INCOME ($ MILLIONS)
Omnicom Group	New York	4154
WPP Group	London	3647
Interpublic Group of Cos.	New York	3385
Dentsu	Tokyo	1988
Young & Rubicam	New York	1498
True North Communications	Chicago	1212
Grey Advertising	New York	1143
Havas Advertising	Paris	1033
Leo Burnett Co.	Chicago	878
Hakuhodo	Tokyo	848
MacManus Group	New York	843
Saatchi & Saatchi	London	657
Publicis Communication	Paris	625
Cordiant Communications Group	London	597
Carlson Marketing Group	Minneapolis	285
TMP Worldwide	New York	274
Asatsu	Tokyo	263
Tokyu Agency	Tokyo	205
Daiko Advertising	Tokyo	204
Abbott Mead Vickers	London	187

SOLUTION

There are 20 advertising organizations, $n = 20$. Q_1 is found by

$$i = \frac{1}{4}(20) = 5$$

Because i is a whole number, Q_1 is found to be the average of the fifth and sixth values from the bottom.

$$Q_1 = \frac{274 + 285}{2} = 279.5$$

Q_2 = median; as there are 20 terms, the median is the average of the tenth and eleventh terms.

$$Q_2 = \frac{843 + 848}{2} = 845.5$$

Q_3 is solved by

$$i = \frac{3}{4}(20) = 15$$

Q_3 is found by averaging the fifteenth and sixteenth terms.

$$Q_3 = \frac{1212 + 1498}{2} = 1355$$

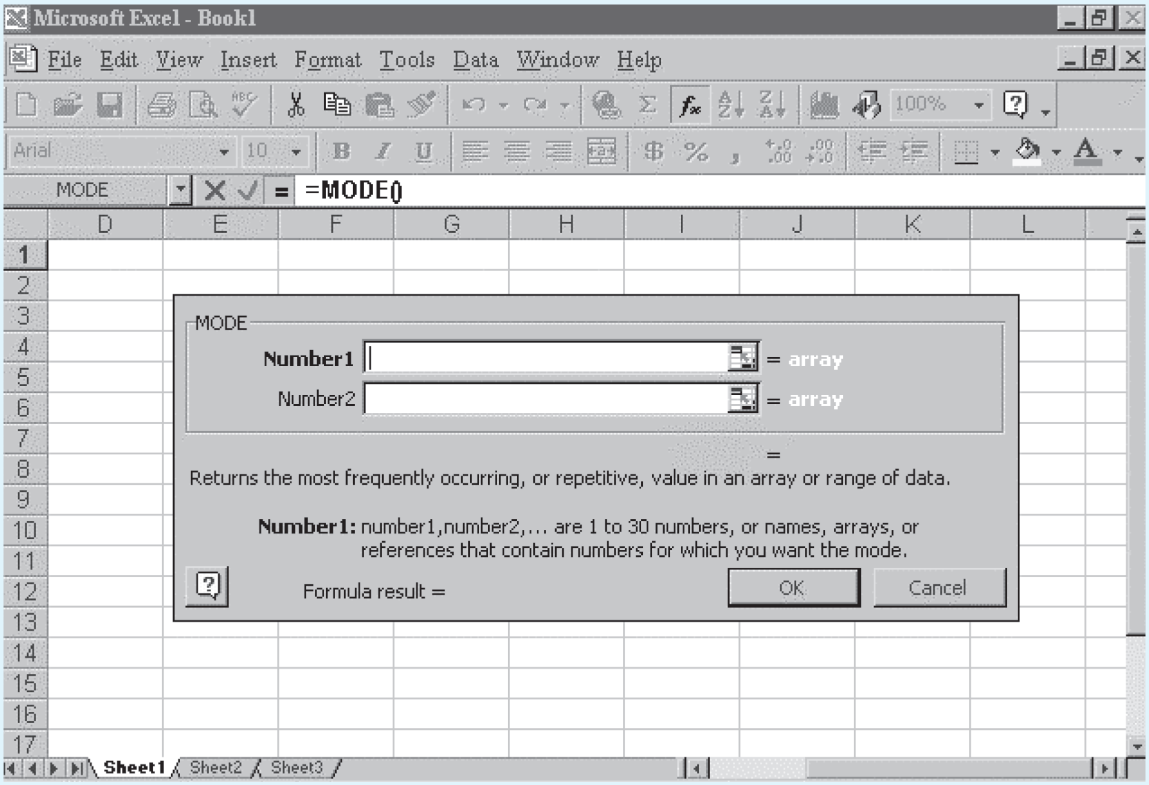
Analysis Using Excel

Excel can compute a mode, a median, a mean, and quartiles. Each of these statistics is accessed using the paste function, f_x. Select **Statistical** from the options presented on the left side of the paste function dialog box, and a long list of statistical options are displayed on the right side. Among the options shown on the right side are **MODE, MEDIAN, AVERAGE** (used to compute means), and **QUARTILE.** The Excel dialog boxes for these four statistics are displayed in Figures 3.2 through 3.5.

To compute a mode, a median, or a mean (average), enter the location of the data in the first box of the dialog box labeled **Number1.** The answer will be displayed on the dialog box and will be shown on the spreadsheet after clicking **OK.** The quartile dialog box also requires that the location of the data be entered in the first box, but this box is labeled **Array** for quartile computation. In the second box of the quartile dialog box labeled **Quart,** insert the number 1 to compute the first quartile, the number 2 to compute the second quartile, and the number 3 to compute the third quartile.

Figure 3.6 displays the Excel output of the mean, median, mode, Q_1, Q_2, and Q_3 for Demonstration Problem 3.1. The answers obtained for the mode, median, mean, and Q_2 are the same as those computed manually in this text. However, Excel defines the first quartile, Q_1, as the $\left[\frac{n+3}{4}\right]^{th}$ item and the third quartile, Q_3, as the $\left[\frac{3n+1}{4}\right]^{th}$ item. Thus, the answers for Q_1 and Q_3 will either be the same or will differ by 1 at the most from the values obtained using methods presented in this chapter.

Figure 3.2 Dialog box for **MODE**



MODE

Number1 = array

Number2 = array

=

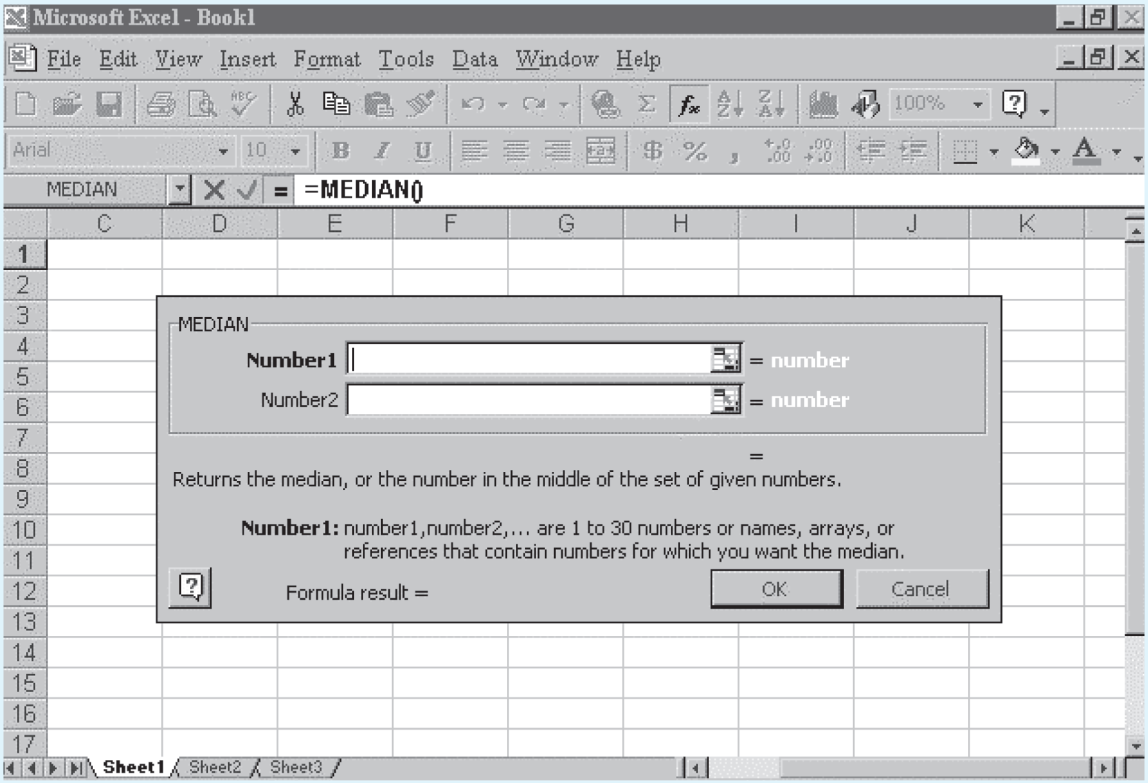
Returns the most frequently occurring, or repetitive, value in an array or range of data.

Number1: number1,number2,... are 1 to 30 numbers, or names, arrays, or references that contain numbers for which you want the mode.

Formula result =

OK Cancel

Figure 3.3 Dialog box for **MEDIAN**



MEDIAN

Number1 = number

Number2 = number

=

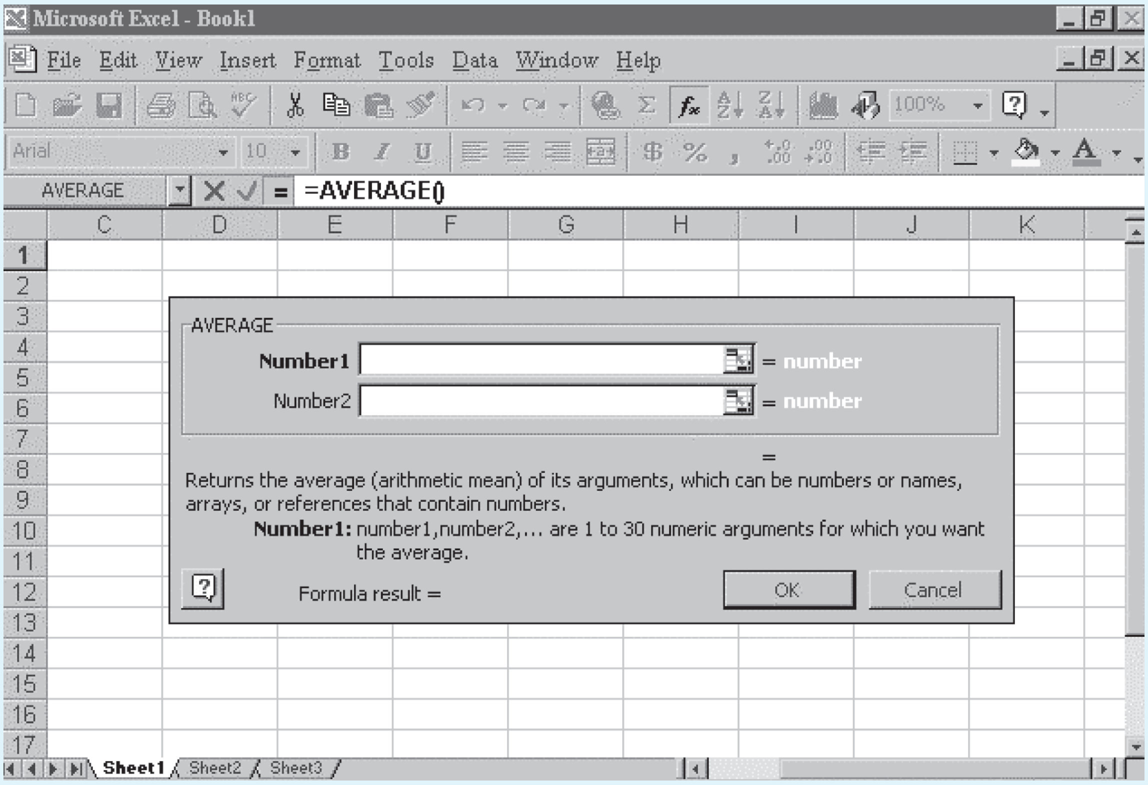
Returns the median, or the number in the middle of the set of given numbers.

Number1: number1,number2,... are 1 to 30 numbers or names, arrays, or references that contain numbers for which you want the median.

Formula result =

OK Cancel

Figure 3.4 Dialog box for **MEAN**



AVERAGE =AVERAGE()

AVERAGE

Number1 = number

Number2 = number

=

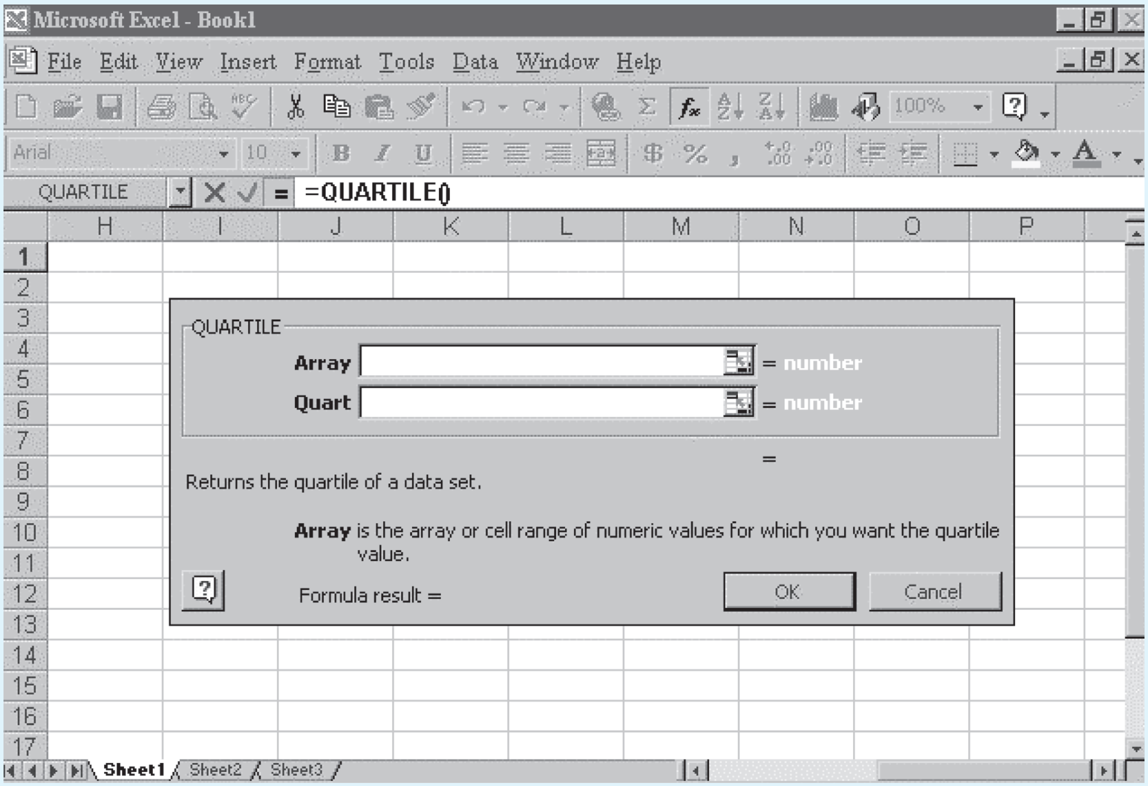
Returns the average (arithmetic mean) of its arguments, which can be numbers or names, arrays, or references that contain numbers.

Number1: number1,number2,... are 1 to 30 numeric arguments for which you want the average.

Formula result =

OK Cancel

Figure 3.5 Dialog box for **QUARTILES**



QUARTILE =QUARTILE()

QUARTILE

Array = number

Quart = number

=

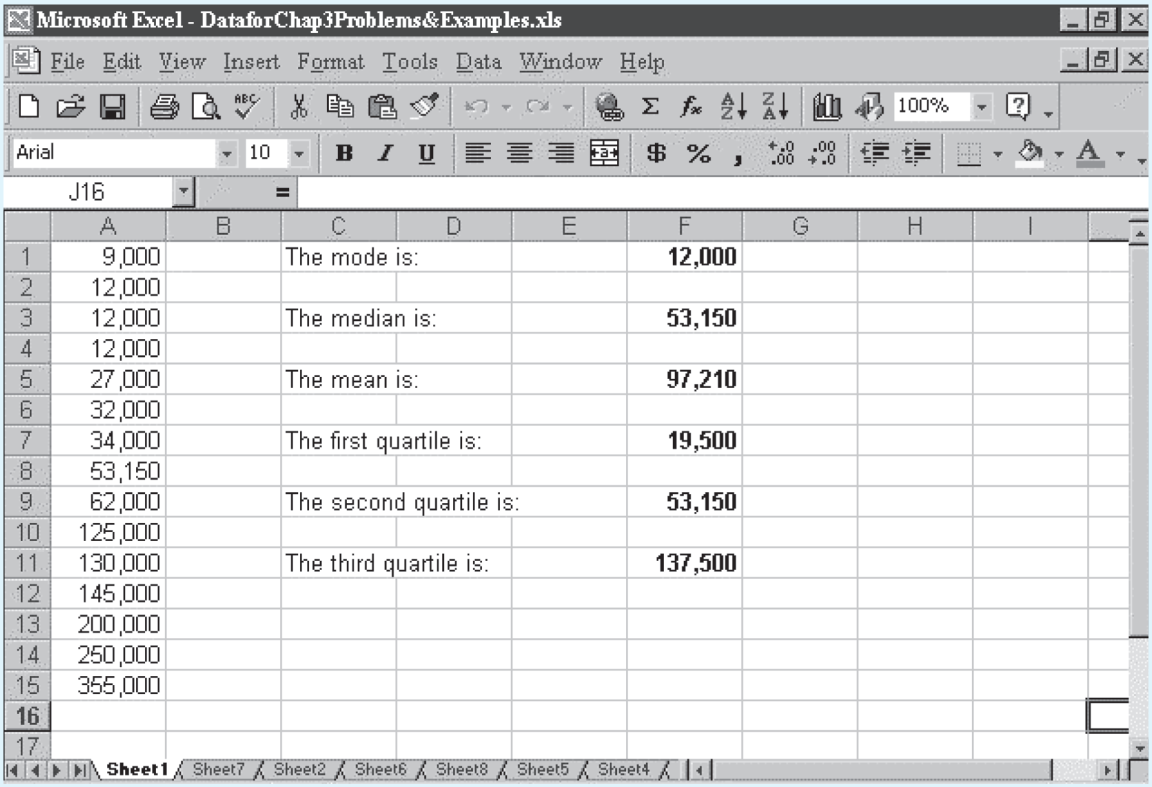
Returns the quartile of a data set.

Array is the array or cell range of numeric values for which you want the quartile value.

Formula result =

OK Cancel

Figure 3.6 Excel output for Demonstration Problem 3.1


Microsoft Excel - DataforChap3Problems&Examples.xls
File Edit View Insert Format Tools Data Window Help
100%
Arial
10
J16
=
Sheet1
Sheet7
Sheet2
Sheet6
Sheet8
Sheet5
Sheet4

	A	B	C	D	E	F	G	H	I
1	9,000		The mode is:			**12,000**			
2	12,000								
3	12,000		The median is:			**53,150**			
4	12,000								
5	27,000		The mean is:			**97,210**			
6	32,000								
7	34,000		The first quartile is:			**19,500**			
8	53,150								
9	62,000		The second quartile is:			**53,150**			
10	125,000								
11	130,000		The third quartile is:			**137,500**			
12	145,000								
13	200,000								
14	250,000								
15	355,000								
16									
17									

3.1 Problems

3.1 Determine the mode for the following numbers.

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

3.2 Determine the median for the numbers in Problem 3. 1.

3.3 Determine the median for the following numbers.

213 345 609 073 167 243 444 524 199 682

3.4 Compute the mean for the following numbers.

17.3 44.5 31.6 40.0 52.8 38.8 30.1 78.5

3.5 Compute the mean for the following numbers.

7 −2 5 9 0 −3 −6 −7 −4 −5 2 −8

3.6 Compute Q_1, Q_2, and Q_3 for the following data.

16 28 29 13 17 20 11 34 32 27 25 30 19 18 33

3.7 Compute Q_1, Q_2, and Q_3 for the following data.

120	138	97	118	172	144
138	107	94	119	139	145
162	127	112	150	143	80
105	116	142	128	116	171

3.8 Shown here are the projected number of cars and light trucks for the year 2000 for the largest automakers in the world, as reported by AutoFacts, a unit of Coopers & Lybrand Consulting. Compute the mean and median. Which of these two measures do you think is most appropriate for summarizing these data and why? What is the value of Q_1, Q_2, and Q_3?

AUTOMAKER	PRODUCTION (THOUSANDS)
General Motors	7880
Ford Motors	6359
Toyota	4580
Volkswagen	4161
Chrysler	2968
Nissan	2646
Honda	2436
Fiat	2264
Peugeot	1767
Renault	1567
Mitsubishi	1535
Hyundai	1434
BMW	1341
Daimler-Benz	1227
Daewoo	898

3.9 The following lists the biggest banks in the world ranked by assets according to *The Banker,* bank reports. Compute the median Q_1 and Q_3.

BANK	ASSETS
BNP-SG-Paribas	$1096
Deutsche-Bank-BT	800
UBS	751
Citigroup	701
Bank of Tokyo-Mitsubishi	653
BankAmerica	595
Credit Suisse	516
Industrial and Commercial Bank of China	489
HSBC	483
Sumitomo Bank	468

3.10 The following lists the number of fatal accidents by scheduled commercial airlines over a 17-year period according to the Air Transport Association of America. Using these data, compute the mean, median, and mode. What is the value of the third quartile?

4 4 4 1 4 2 4 3 8 6 4 4 1 4 2 3 3

3.2 Measures of Variability

Measures of variability
Statistics that describe the spread or dispersion of a set of data.

Measures of central tendency yield information about particular points of a data set. However, researchers can use another group of analytic tools to describe a set of data. These tools are **measures or variability,** which *describe the spread or the dispersion of a set of data.* Using measures of variability in conjunction with measures of central tendency makes possible a more complete numerical description of the data.

For example, a company has 25 salespeople in the field, and the median annual sales figure for these people is $1,200,000. Are the salespeople being successful as a *group* or not? The median provides information about the sales of the person in the middle, but what about the other salespeople? Are all of them selling $ 1,200,000 annually, or do the

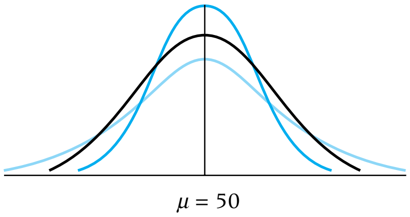


Figure 3.7

Three distributions with the same mean but different dispersions

sales figures vary widely, with one person selling $5,000,000 annually and another selling only $150,000 annually? Measures of variability provide the additional information necessary to answer that question.

Figure 3.7 shows three distributions in which the mean of each distribution is the same ($\mu = 50$) but the variabilities differ. Observation of these distributions shows that a measure of variability is necessary to complement the mean value in describing the data. This section focuses on seven measures of variability: range, interquartile range, mean absolute deviation, variance, standard deviation, *Z* scores, and coefficient of variation.

Range

Range
The difference between the largest and the smallest values in a set of numbers.

The **range** is *the difference between the largest value of a data set and the smallest value.* Although it is usually a single numeric value, some researchers define the range as *the ordered pair of smallest and largest numbers (smallest, largest).* It is a crude measure of variability, describing the distance to the outer bounds of the data set. It reflects those extreme values because it is constructed from them. An advantage of the range is its ease of computation. One important use of the range is in quality assurance, where the range is used to construct control charts. A disadvantage of the range is that because it is computed with the values that are on the extremes of the data it is affected by extreme values and therefore its application as a measure of variability is limited.

The data in Table 3.1 represent the offer prices for the 20 largest U.S. initial public offerings in a recent year. The lowest offer price was $7.00 and the highest price was $43.25. The range of the offer prices can be computed as the difference of the highest and lowest values:

$$\text{Range} = \text{Highest} - \text{Lowest} = \$43.25 - \$7.00 = \$36.25$$

Interquartile Range

Interquartile range
The range of values between the first and the third quartile.

Another measure of variability is the **interquartile range.** The interquartile range is *the range of values between the first and third quartile.* Essentially, it is the range of the middle 50% of the data, and it is determined by computing the value of $Q_3 - Q_1$. The interquartile range is especially useful in situations where data users are more interested in values toward the middle and less interested in extremes. In describing a real estate housing market, realtors might use the interquartile range as a measure of housing prices when describing the middle half of the market when buyers are interested in houses in the midrange. In addition, the interquartile range is used in the construction of box and whisker plots.

$$Q_3 - Q_1$$

INTERQUARTILE RANGE

The following lists the top 15 trading partners of the United States by U.S. exports to the country in a recent year according to the U.S. Census Bureau.

COUNTRY	EXPORTS ($ BILLIONS)
Canada	$151.8
Mexico	71.4
Japan	65.5
United Kingdom	36.4
South Korea	25.0
Germany	24.5
Taiwan	20.4
Netherlands	19.8
Singapore	17.7
France	16.0
Brazil	15.9
Hong Kong	15.1
Belgium	13.4
China	12.9
Australia	12.1

What is the interquartile range for these data? The process begins by computing the first and third quartiles as follows.

Solving for Q_1 when $n = 15$:

$$i = \frac{1}{4}(15) = 3.75$$

Since i is not a whole number, Q_1 is found as the 4th term from the bottom.

$$Q_1 = 15.1$$

Solving for Q_3:

$$i = \frac{3}{4}(15) = 11.25$$

Since i is not a whole number, Q_3 is found as the 12th term from the bottom.

$$Q_3 = 36.4$$

The interquartile range is:

$$Q_3 - Q_1 = 36.4 - 15.1 = 21.3$$

The middle 50% of the exports for the top 15 United States trading partners spans a range of 21.3 ($ billions).

Mean Absolute Deviation, Variance, and Standard Deviation

Three other measures of variability are the variance, the standard deviation, and the mean absolute deviation. They are obtained through similar processes and are therefore presented together. These measures are not meaningful unless the data are at least interval-level data. The variance and standard deviation are widely used in statistics. Although the standard deviation has some stand-alone potential, the importance of variance and standard deviation lies mainly in their role as tools used in conjunction with other statistical devices.

Suppose a small company has started a production line to build computers. During the first five weeks of production, the output is 5, 9, 16, 17, and 18 computers, respectively. Which descriptive statistics could the owner use to measure the early progress of production? In an attempt to summarize these figures, he could compute a mean.

$$\begin{array}{c} X \\ \hline 5 \\ 9 \\ 16 \\ 17 \\ 18 \\ \hline \end{array}$$

$$\Sigma X = 65 \qquad \mu = \frac{\Sigma X}{N} = \frac{65}{5} = 13$$

What is the variability in these five weeks of data? One way for the owner to begin to look at the spread of the data is to subtract the mean from each data value. *Subtracting the mean from each value of data* yields the **deviation from the mean** $(X - \mu)$. Table 3.2 shows these deviations for the computer company production. Note that some deviations from the mean are positive and some are negative. Figure 3.8 shows that geometrically the negative deviations represent values that are below (to the left of) the mean and positive deviations represent values that are above (to the right of) the mean.

Deviation from the mean The difference between a number and the average of the set of numbers of which the number is a part.

An examination of deviations from the mean can reveal information about the variability of data. However, the deviations are used mostly as a tool to compute other measures of variability. Note that in both Table 3.2 and Figure 3.8 these deviations total zero. This phenomenon applies to all cases. For a given set of data, the sum of all deviations from the arithmetic mean is always zero.

$$\Sigma(X - \mu) = 0$$

Sum of Deviations from the Arithmetic Mean Is Always Zero

TABLE 3.2
Deviations from the Mean for Computer Production

NUMBER (X)	DEVIATIONS FROM THE MEAN $(X - \mu)$
5	$5 - 13 = -8$
9	$9 - 13 = -4$
16	$16 - 13 = +3$
17	$17 - 13 = +4$
18	$18 - 13 = +5$
$\Sigma X = 65$	$\Sigma(X - \mu) = 0$

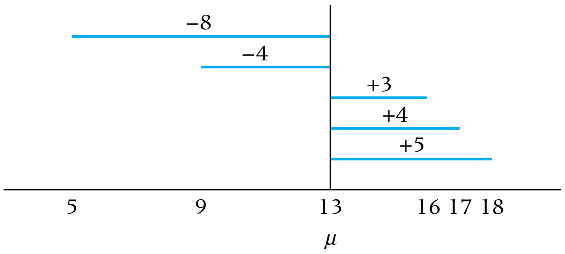


Figure 3.8

Geometric distances from the mean (from Table 3.2)

This property requires considering alternative ways to obtain measures of variability.

One obvious way to force the sum of deviations to have a nonzero total is to take the absolute value of each deviation around the mean. Utilizing the absolute value of the deviations about the mean makes solving for the mean absolute deviation possible.

Mean absolute deviation (MAD)
The average of the absolute values of the deviations around the mean for a set of numbers.

Mean Absolute Deviation

The **mean absolute deviation (MAD)** is *the average of the absolute values of the deviations around the mean for a set of numbers.*

MEAN ABSOLUTE DEVIATION

$$\text{MAD} = \frac{\Sigma|X - \mu|}{N}$$

Using the data from Table 3.2, the computer company owner can compute a mean absolute deviation by taking the absolute values of the deviations and averaging them, as shown in Table 3.3. The mean absolute deviation for the computer production data is 4.8.

Because it is computed by using absolute values, the mean absolute deviation is less useful in statistics than other measures of dispersion. However, in the field of forecasting, it is used occasionally as a measure of error.

Variance

Variance
The average of the squared deviations about the arithmetic mean for a set of numbers.

Because absolute values are not conducive to easy manipulation, mathematicians developed an alternative mechanism for overcoming the zero-sum property of deviations from the mean. This approach utilizes the square of the deviations from the mean. The result is the variance, an important measure of variability.

The **variance** is *the average of the squared deviations about the arithmetic mean for a set of numbers.* The population variance is denoted by σ^2.

POPULATION VARIANCE

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

Table 3.4 shows the original production numbers for the computer company, the deviations from the mean, and the squared deviations from the mean.

Sum of squares of *X*
The sum of the squared deviations about the mean of a set of values.

The sum of the squared deviations about the mean of a set of values—called the **sum of squares of *X*** and sometimes abbreviated as SS_X—is used throughout statistics. For the computer company, this value is 130. Dividing it by the number of data values (5 wk) yields the variance for computer production.

$$\sigma^2 = \frac{130}{5} = 26.0$$

Because the variance is computed from squared deviations, the final result is expressed in terms of squared units of measurement. Statistics measured in squared units are problematic to interpret. Consider, for example, Mattel Toys attempting to interpret production costs in terms of squared dollars or Troy-Built measuring production

TABLE 3.3
MAD for Computer Production Data

X	$X - \mu$	$\lvert X - \mu \rvert$
5	−8	+8
9	−4	+4
16	+3	+3
17	+4	+4
18	+5	+5
$\Sigma X = 65$	$\Sigma(X - \mu) = 0$	$\Sigma \lvert X - \mu \rvert = 24$

$$\text{MAD} = \frac{\Sigma \lvert X - \mu \rvert}{N} = \frac{24}{5} = 4.8$$

TABLE 3.4
Computing a Variance and a Standard Deviation from the Computer Production Data

X	$X - \mu$	$(X - \mu)^2$
5	−8	64
9	−4	16
16	+3	9
17	+4	16
18	+5	25
$\Sigma X = 65$	$\Sigma(X - \mu) = 0$	$\Sigma(X - \mu)^2 = 130$

$$\text{SS}_X = \Sigma(X - \mu)^2 = 130$$

$$\text{Variance} = \sigma^2 = \frac{SS_X}{N} = \frac{\Sigma(X - \mu)^2}{N} = \frac{130}{5} = 26.0$$

$$\text{Standard deviation} = \sigma = \sqrt{\frac{\Sigma(X - \mu)^2}{N}} = \sqrt{\frac{130}{5}} = 5.1$$

output variation in terms of squared lawn mowers. Therefore, when used as a descriptive measure, variance can be considered as an intermediate calculation in the process of obtaining the sample standard deviation.

Standard Deviation

The standard deviation is a popular measure of variability. It is used both as a separate entity and as a part of other analyses, such as computing confidence intervals and in hypothesis testing (see Chapters 8, 9, and 10).

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

Population Standard Deviation

The **standard deviation** is *the square root of the variance.* The population standard deviation is denoted by σ.

Standard deviation
The square root of the variance.

Like the variance, the standard deviation utilizes the sum of the squared deviations about the mean (SS_X). It is computed by averaging these squared deviations (SS_X/N) and taking the square root of that average. One feature of the standard deviation that distinguishes it from a variance is that the standard deviation is expressed in the same units as the raw data, whereas the variance is expressed in those units squared. Table 3.4 shows the standard deviation for the computer production company: $\sqrt{26}$, or 5.1.

What does a standard deviation of 5.1 mean? The meaning of standard deviation is more readily understood from its use, which is explored in the next section. Although the standard deviation and the variance are closely related and can be computed from each other, differentiating between them is important, because both are widely used in statistics.

Meaning of Standard Deviation

What is a standard deviation? What does it do, and what does it mean? There is no precise way of defining a standard deviation other than reciting the formula used to compute it. However, insight into the concept of standard deviation can be gleaned by viewing the manner in which it is applied. One way of applying the standard deviation is the empirical rule.

Empirical rule
A guideline that states the approximate percentage of values that fall within a given number of standard deviations of a mean of a set of data that are normally distributed.

EMPIRICAL RULE The **empirical rule** is a very important rule of thumb that is used to state the approximate percentage of values that lie within a given number of standard deviations from the mean of a set of data if the data are normally distributed.

The empirical rule is used only for three numbers of standard deviations: 1σ, 2σ, and 3σ. More detailed analysis of other numbers of σ values is presented in Chapter 6. Also discussed in further detail in Chapter 6 is the normal distribution, a unimodal, symmetrical distribution that is bell (or mound) shaped. The requirement that the data be normally distributed contains some tolerance, and the empirical rule generally applies so long as the data are approximately mound shaped.

EMPIRICAL RULE*

DISTANCE FROM THE MEAN	VALUES WITHIN DISTANCE
$\mu \pm 1\sigma$	68%
$\mu \pm 2\sigma$	95%
$\mu \pm 3\sigma$	99.7%

*Based on the assumption that the data are approximately normally distributed.

If a set of data is normally distributed, or bell shaped, approximately 68% of the data values are within one standard deviation of the mean, 95% are within two standard deviations, and almost 100% are within three standard deviations.

For example, suppose a recent report states that for California the average statewide price of a gallon of regular gasoline is \$1.52. Suppose regular gasoline prices vary across the state with a standard deviation of \$0.08 and are normally distributed. According to the empirical rule, approximately 68% of the prices should fall within $\mu \pm 1\sigma$, or \$1.52 ± 1 (\$0.08). Approximately 68% of the prices would be between \$1.44 and \$1.60, as shown in Figure 3.9A. Approximately 95% should fall within $\mu \pm 2\sigma$ or \$1.52 ± 2 (\$0.08) = \$1.52 ± \$0.16, or between \$1.36 and \$1.68, as shown in Figure 3.9B. Nearly all regular gasoline prices (99.7%) should fall between \$1.28 and \$1.76 ($\mu \pm 3\sigma$).

Note that since 68% of the gasoline prices lie within one standard deviation of the mean, approximately 32% are outside this range. Since the normal distribution is symmetrical, the 32% can be split in half such that 16% lie in each tail of the distribution. Thus, approximately 16% of the gasoline prices should be less than \$1.44 and approximately 16% of the prices should be greater than \$1.60.

Because many phenomena are distributed approximately in a bell shape, including most human characteristics, such as height and weight, the empirical rule applies in many situations and is widely used.

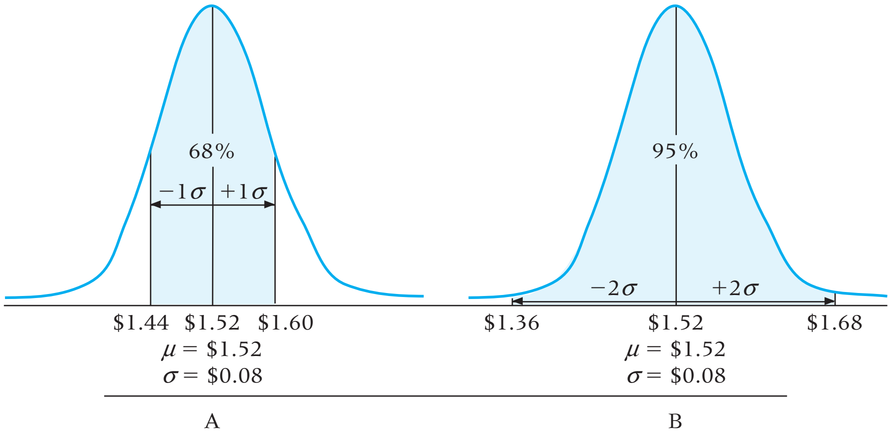


Figure 3.9

Empirical rule for one and two standard deviations of gasoline prices

DEMONSTRATION PROBLEM 3.3

A company produces a lightweight valve that is specified to weigh 1365 g. Unfortunately, because of imperfections in the manufacturing process not all of the valves produced weigh exactly 1365 grams. In fact, the weights of the valves produced are normally distributed with a mean weight of 1365 grams and a standard deviation of 294 grams. Within what range of weights would approximately 95% of the valve weights fall? Approximately 16% of the weights would be more than what value? Approximately 0.15% of the weights would be less than what value?

SOLUTION

Since the valve weights are normally distributed, the empirical rule applies. According to the empirical rule, approximately 95% of the weights should fall within $\mu \pm 2\sigma = 1365 \pm 2(294) = 1365 \pm 588$. Thus, approximately 95% should fall between 777 and 1953. Approximately 68% of the weights should fall within $\mu \pm 1\sigma$ and 32% should fall outside this interval. Because the normal distribution is symmetrical, approximately 16% should lie above $\mu + 1\sigma = 1365 + 294 = 1659$. Approximately 99.7% of the weights should fall within $\mu \pm 3\sigma$ and .3% should fall outside this interval. Half of these or .15% should lie below $\mu - 3\sigma = 1365 - 3(294) = 1365 - 882 = 483$.

Population versus Sample Variance and Standard Deviation

The sample variance is denoted by S^2 and the sample standard deviation by S. Computation of the sample variance and standard deviation differs slightly from computation of the population variance and standard deviation. The main use for sample variances and standard deviations is as estimators of population variances and standard deviations. Using $n - 1$ in the denominator of a sample variance or standard deviation, rather than n, results in a better estimate of the population values.

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

SAMPLE VARIANCE

$$S = \sqrt{S^2}$$

SAMPLE STANDARD DEVIATION

Shown here is a sample of six of the largest accounting firms in the United States and the number of partners associated with each firm as reported by the *Public Accounting Report.*

FIRM	NUMBER OF PARTNERS
Price Waterhouse	1062
McGladrey & Pullen	381
Deloitte & Touche	1719
Andersen Worldwide	1673
Coopers & Lybrand	1277
BDO Seidman	217

The sample variance and sample standard deviation can be computed by:

X	$(X-\bar{X})^2$
1062	51.41
381	454,046.87
1719	441,121.79
1673	382,134.15
1277	49,359.51
217	701,959.11
$\Sigma X = 6329$	$SS_X = \Sigma(X-\bar{X})^2 = 2{,}028{,}672.84$

$$\bar{X} = \frac{6329}{6} = 1054.83$$

$$S^2 = \frac{\Sigma(X-\bar{X})^2}{n-1} = \frac{2{,}028{,}627.84}{5} = 405{,}734.57$$

$$S = \sqrt{S^2} = \sqrt{405{,}734.57} = 636.97$$

The sample variance is 405,734.57 and the sample standard deviation is 636.97.

Computational Formulas for Variance and Standard Deviation

An alternative method of computing variance and standard deviation, sometimes referred to as the computational method or shortcut method, is available. Algebraically,

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

and

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

Substituting these equivalent expressions into the original formulas for variance and standard deviation yields the following computational formulas.

COMPUTATIONAL FORMULA FOR POPULATION VARIANCE AND STANDARD DEVIATION

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

$$\sigma = \sqrt{\sigma^2}$$

TABLE 3.5 Computational Formula Calculations of Variance and Standard Deviation for Computer Production Data

X	X^2
5	25
9	81
16	256
17	289
18	324
$\Sigma X = 65$	$\Sigma X^2 = 975$

$$\sigma^2 = \frac{975 - \frac{(65)^2}{5}}{5} = \frac{975 - 845}{5} = \frac{130}{5} = 26$$

$$\sigma = \sqrt{26} = 5.1$$

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

$$S = \sqrt{S^2}$$

Computational Formula for Sample Variance and Standard Deviation

These computational formulas utilize the sum of the X values and the sum of the X^2 values instead of the difference between the mean and each value and computed deviations. In the pre-calculator/computer era, this method usually was faster and easier than using the original formulas.

For situations in which the mean is already computed or is given, alternative forms of these formulas are

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

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

Using the computational method, the owner of the start-up computer production company can compute a population variance and standard deviation for the production data, as shown in Table 3.5. (Compare these results with those in Table 3.4.)

DEMONSTRATION PROBLEM 3.4

The effectiveness of district attorneys can be measured by several variables, including the number of convictions per month, the number of cases handled per month, and the total number of years of conviction per month. A researcher uses a sample of five district attorneys in a city. She determines the total number of years of conviction that each attorney won against defendants during the past month, as reported in the first column in the following tabulations. Compute the mean absolute deviation, the variance, and the standard deviation for these figures.

SOLUTION

The researcher computes the mean absolute deviation, the variance, and the standard deviation for these data in the following manner.

X	$\lvert X-\bar{X}\rvert$	$(X-\bar{X})^2$
55	41	1,681
100	4	16
125	29	841
140	44	1,936
60	36	1,296
$\Sigma X = 480$	$\Sigma\lvert X-\bar{X}\rvert = 154$	$SS_X = 5{,}770$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{480}{5} = 96$$

$$\text{MAD} = \frac{154}{5} = 30.8$$

$$S^2 = \frac{5{,}770}{4} = 1{,}442.5 \quad \text{and} \quad S = \sqrt{S^2} = 37.98$$

She then uses computational formulas to solve for S^2 and S and compares the results.

X	X^2
55	3,025
100	10,000
125	15,625
140	19,600
60	3,600
$\Sigma X = 480$	$\Sigma X^2 = 51{,}850$

$$S^2 = \frac{51{,}850 - \frac{(480)^2}{5}}{4} = \frac{51{,}850 - 46{,}080}{4} = \frac{5{,}770}{4} = 1{,}442.5$$

$$S = \sqrt{1{,}442.5} = 37.98$$

The results are the same. The sample standard deviation obtained by both methods is 37.98, or 38, years.

***Z* score**
The number of standard deviations a value (X) is above or below the mean of a set of numbers when the data are normally distributed.

Z Scores

A ***Z* score** represents *the number of standard deviations a value (X) is above or below the mean of a set of numbers when the data are normally distributed.* Using Z scores allows translation of a value's raw distance from the mean into units of standard deviations.

Z SCORE

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

For samples,

$$Z = \frac{X-\bar{X}}{S}.$$

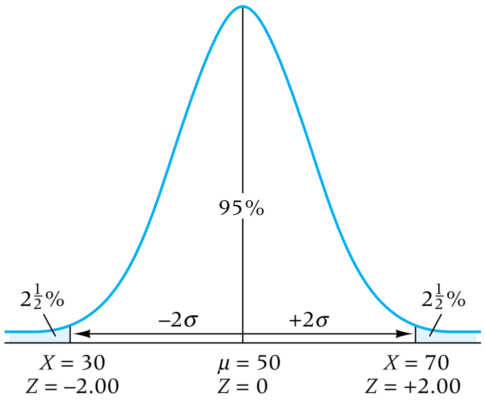


Figure 3.10

Percentage breakdown of scores two standard deviations from the mean

If a Z score is negative, the raw value (X) is below the mean. If the Z score is positive, the raw value (X) is above the mean.

For example, for a data set that is normally distributed with a mean of 50 and a standard deviation of 10, suppose a statistician wants to determine the Z score for a value of 70. This value ($X = 70$) is 20 units above the mean, so the Z value is

$$Z = \frac{70 - 50}{10} = +\frac{20}{10} = +2.00$$

This Z score signifies that the raw score of 70 is two standard deviations above the mean. How is this Z score interpreted? The empirical rule states that 95% of all values are within two standard deviations of the mean if the data are approximately normally distributed. Figure 3.10 shows that because the value of 70 is two standard deviations above the mean ($Z = +\ 2.00$), 95% of the values are between 70 and the value ($X = 30$), that is two standard deviations below the mean ($Z = \frac{30 - 50}{10} = -\ 2.00$). As 5% of the values are outside the range of two standard deviations from the mean and the normal distribution is symmetrical, 2½% (½ of the 5%) are below the value of 30. Thus 97½% of the values are below the value of 70. Because a Z score is the number of standard deviations an individual data value is from the mean, the empirical rule can be restated in terms of Z scores.

Between $Z = -\ 1.00$ and $Z = +\ 1.00$ are approximately 68% of the values.

Between $Z = -\ 2.00$ and $Z = +\ 2.00$ are approximately 95% of the values.

Between $Z = -3.00$ and $Z = +\ 3.00$ are approximately 99.7% of the values.

The topic of Z scores is discussed more extensively in Chapter 6.

Coefficient of Variation

The **coefficient of variation** is a statistic that is *the ratio of the standard deviation to the mean expressed in percentage* and is denoted CV.

Coefficient of variation (CV)
The ratio of the standard deviation to the mean, expressed as a percentage.

COEFFICIENT OF VARIATION

$$\text{CV} = \frac{\sigma}{\mu}(100)$$

For sample data, $\text{CV} = \dfrac{S}{\bar{X}}(100)$.

The coefficient of variation essentially is a relative comparison of a standard deviation to its mean. The coefficient of variation can be useful in comparing standard deviations that have been computed from data with different means.

Suppose five weeks of average prices for stock A are 57, 68, 64, 71, and 62. To compute a coefficient of variation for these prices, first determine the mean and standard deviation: $\mu = 64.40$ and $\sigma = 4.84$. The coefficient of variation is:

$$\text{CV}_A = \frac{\sigma_A}{\mu_A}(100) = \frac{4.84}{64.40}(100) = .075 = 7.5\%$$

The standard deviation is 7.5% of the mean.

Sometimes financial investors use the coefficient of variation or the standard deviation or both as measures of risk. Imagine a stock with a price that never changes. There is no risk of losing money from the price going down because there is no variability to the price. Suppose, in contrast, that the price of the stock fluctuates wildly. An investor who buys at a low price and sells for a high price can make a nice profit. However, if the price drops below what the investor buys it for, there is a potential for loss. The greater the variability, the more the potential for loss. Hence, investors use measures of variability such as standard deviation or coefficient of variation to determine the risk of a stock. What does the coefficient of variation tell us about the risk of a stock that the standard deviation does not?

Suppose the average prices for a second stock, B, over these same five weeks are 12, 17, 8, 15, and 13. The mean for stock B is 13.00 with a standard deviation of 3.03. The coefficient of variation can be computed for stock B as:

$$\text{CV}_B = \frac{\sigma_B}{\mu_B}(100) = \frac{3.03}{13}(100) = .233 = 23.3\%$$

The standard deviation for stock B is 23.3% of the mean.

With the standard deviation as the measure of risk, stock A is more risky over this period of time because it has a larger standard deviation. However, the average price of stock A is almost five times as much as that of stock B. Relative to the amount invested in stock A, the standard deviation of \$4.84 may not represent as much risk as the standard deviation of \$3.03 for stock B, which has an average price of only \$13.00. The coefficient of variation reveals the risk of a stock in terms of the size of standard deviation relative to the size of the mean (in percentage).

Stock B has a coefficient of variation that is nearly three times as much as the coefficient of variation for stock A. Using coefficient of variation as a measure of risk indicates that stock B is riskier.

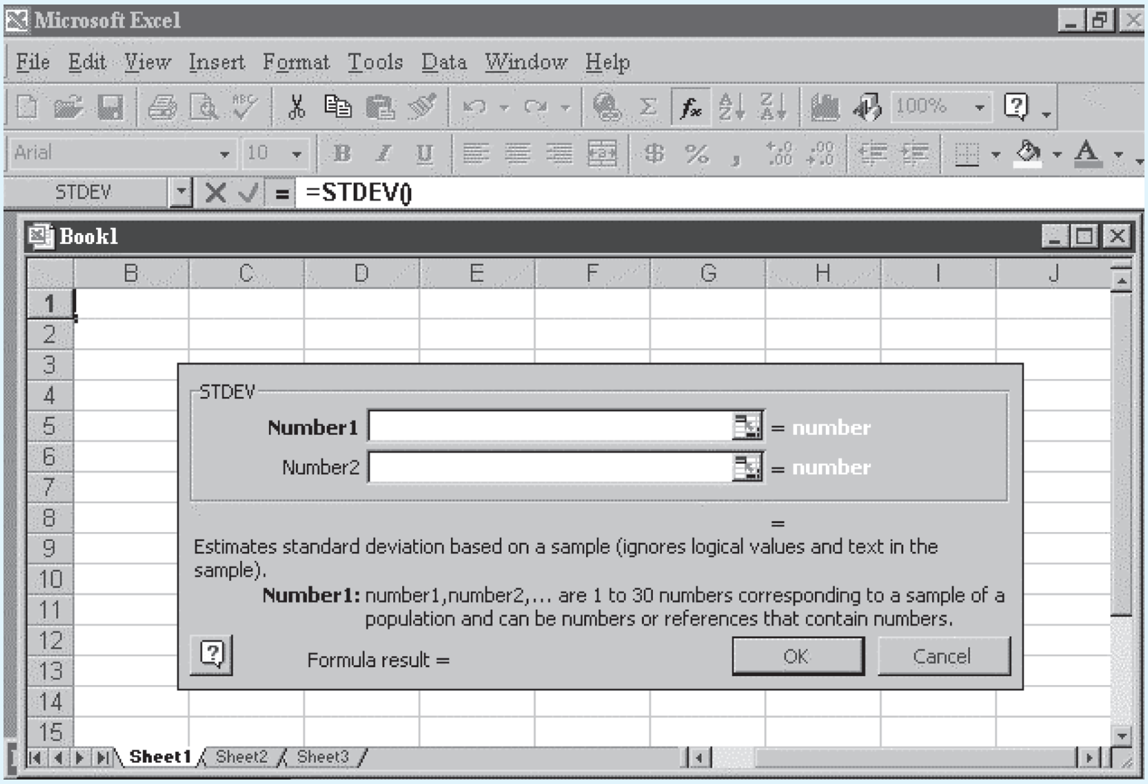
The choice of whether to use a coefficient of variation or raw standard deviations to compare multiple standard deviations is a matter of preference. The coefficient of variation also provides an optional method of interpreting the value of a standard deviation.

Analysis Using Excel

Excel can compute the variance and the standard deviation for both a population and a sample. The range is computed as part of *Summary Statistics,* which are discussed later in Section 3.4. To compute the variance and standard deviation, begin with the paste function ***fx.*** Select **Statistical** from the left side of the paste function dialog box. Included in the menu on the right side of this dialog box is **STDEV,** which computes the sample standard deviation, **STDEVP,** which computes the population standard deviation, **VAR,** which computes the sample variance, and **VARP,** which computes the population variance. The dialog boxes for each of these functions are shown in Figures 3.11, 3.12, 3.13, and 3.14. In each of these dialog boxes, place the location of the data to be analyzed in the line labeled **Number1.** The resulting answer will be displayed on the dialog box; after clicking **OK,** the answer will be displayed on the worksheet.

Figure 3.15 displays the sample standard deviation and sample variance for the attorney data presented in Demonstration Problem 3.4. In addition, Figure 3.15 contains the population standard deviation and population variance for the computer production data presented at the beginning of the section. Note that the answers obtained from Excel are the same as those computed manually in the book.

Figure 3.11 Dialog box for **STDEV**



STDEV =STDEV()

STDEV

Number1 = number

Number2 = number

=

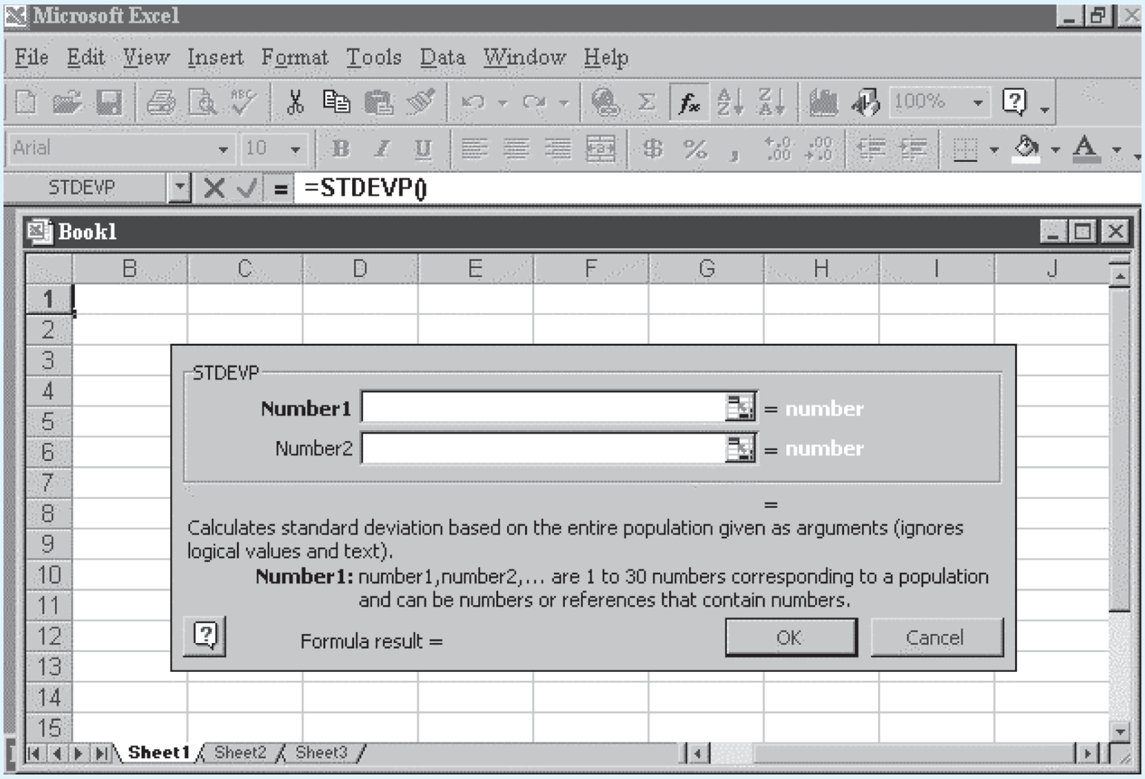
Estimates standard deviation based on a sample (ignores logical values and text in the sample).

Number1: number1,number2,... are 1 to 30 numbers corresponding to a sample of a population and can be numbers or references that contain numbers.

Formula result =

OK Cancel

Figure 3.12 Dialog box for **STDEVP**

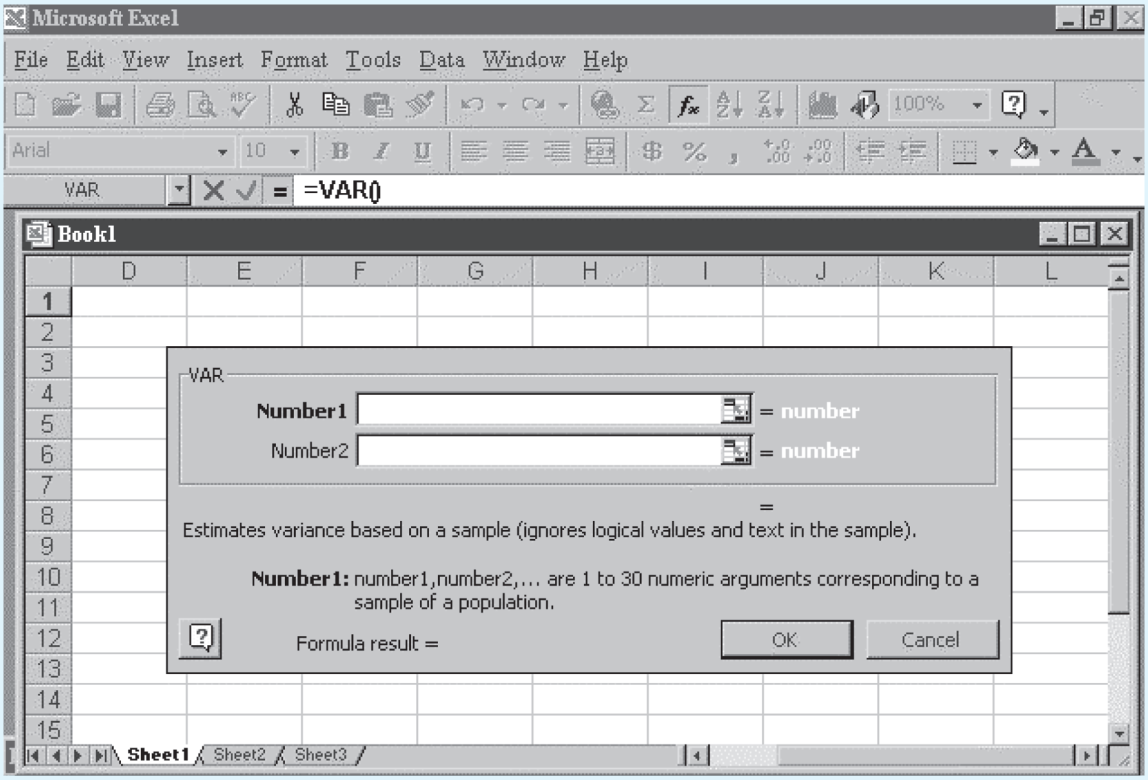


Calculates standard deviation based on the entire population given as arguments (ignores logical values and text).

Number1: number1,number2,... are 1 to 30 numbers corresponding to a population and can be numbers or references that contain numbers.

Formula result =

Figure 3.13 Dialog box for **VAR**

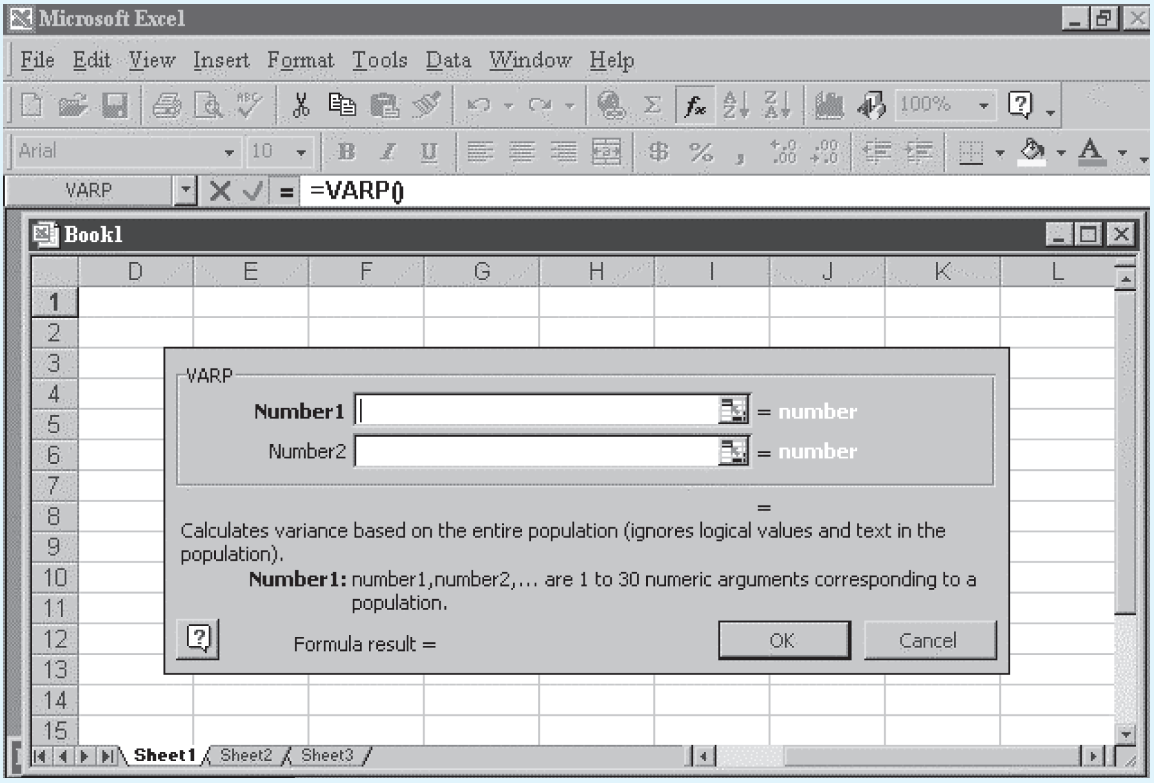


Estimates variance based on a sample (ignores logical values and text in the sample).

Number1: number1,number2,... are 1 to 30 numeric arguments corresponding to a sample of a population.

Formula result =

Figure 3.14 Dialog box for **VARP**



=VARP()

VARP

Number1 = number

Number2 = number

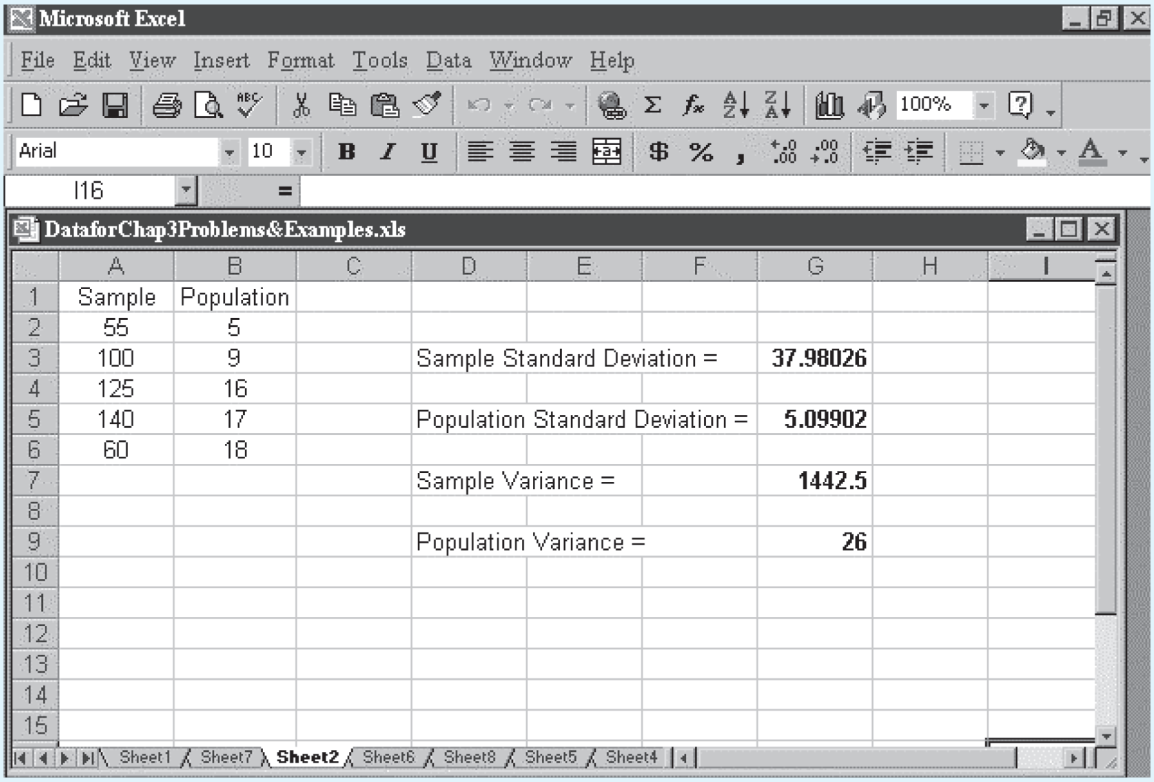
Calculates variance based on the entire population (ignores logical values and text in the population).

Number1: number1,number2,... are 1 to 30 numeric arguments corresponding to a population.

Formula result =

OK Cancel

Figure 3.15 Excel standard deviation and variance output


Microsoft Excel
File Edit View Insert Format Tools Data Window Help
100%
Arial
10
I16
DataforChap3Problems&Examples.xls
Sheet1 Sheet7 Sheet2 Sheet6 Sheet8 Sheet5 Sheet4

	A	B	C	D	E	F	G
1	Sample	Population					
2	55	5					
3	100	9		Sample Standard Deviation =			37.98026
4	125	16					
5	140	17		Population Standard Deviation =			5.09902
6	60	18					
7				Sample Variance =			1442.5
8							
9				Population Variance =			26

3.2 Problems

3.11 A data set contains the following seven values.

6 2 4 9 1 3 5

a. Find the range.
b. Find the mean absolute deviation.
c. Find the population variance.
d. Find the population standard deviation.
e. Find the interquartile range.
f. Find the Z score for each value.

3.12 A data set contains the following eight values.

4 3 0 5 2 9 4 5

a. Find the range.
b. Find the mean absolute deviation.
c. Find the sample variance.
d. Find the sample standard deviation.
e. Find the interquartile range.

3.13 A data set contains the following six values.

12 23 19 26 24 23

a. Find the population standard deviation using the formula containing the mean (the original formula).
b. Find the population standard deviation using the computational formula.
c. Compare the results. Which formula was faster to use? Which formula do you prefer? Why do you think the computational formula is sometimes referred to as the "shortcut" formula?

3.14 Use Excel to find the sample variance and sample standard deviation for the following data.

57	88	68	43	93
63	51	37	77	83
66	60	38	52	28
34	52	60	57	29
92	37	38	17	67

3.15 Use Excel to find the population variance and population standard deviation for the following data.

123	090	546	378
392	280	179	601
572	953	749	075
303	468	531	646

3.16 Determine the interquartile range on the following data.

44	18	39	40	59
46	59	37	15	73
23	19	90	58	35
82	14	38	27	24
71	25	39	84	70

3.17 Compare the variability of the following two sets of data by using both the standard deviation and the coefficient of variation.

DATA SET 1	DATA SET 2
49	159
82	121
77	138
54	152

3.18 A sample of 12 small accounting firms reveals the following numbers of professionals per office.

7	10	9	14	11	8
5	12	8	3	13	6

a. Determine the mean absolute deviation.
b. Determine the variance.
c. Determine the standard deviation.
d. Determine the interquartile range.
e. What is the *Z* score for the firm that has six professionals?
f. What is the coefficient of variation for this sample?

3.19 The following is a list supplied by Marketing Intelligence Service, Ltd., of the companies with the most new products in a recent year.

COMPANY	NUMBER OF NEW PRODUCTS
Avon Products, Inc.	768
L'Oreal	429
Unilever U.S. Inc.	323
Revlon, Inc.	306
Garden Botanika	286
Philip Morris, Inc.	262
Procter & Gamble Co.	215
Nestlé	172
Paradiso Ltd.	162
Tsumura International, Inc.	148
Grand Metropolitan, Inc.	145

a. Find the range.
b. Find the mean absolute deviation.
c. Find the population variance.
d. Find the population standard deviation.
e. Find the interquartile range.
f. Find the *Z* score for Nestlé.
g. Find the coefficient of variation.

3.20 A distribution of numbers is approximately bell-shaped. If the mean of the numbers is 125 and the standard deviation is 12, between what two numbers would approximately 68% of the values be? Between what two numbers would 95% of the values be? Between what two values would 99.7% of the values be?

3.21 The time needed to assemble a particular piece of furniture with experience is normally distributed with a mean time of 43 minutes. If 68% of the assembly times are between 40 and 46 minutes, what is the value of the standard deviation? Suppose 99.7% of the assembly times are between 35 and 51 minutes and the mean is still 43 minutes. What would the value of the standard deviation be now?

3.22 Environmentalists are concerned about emissions of sulfur dioxide into the air. The average number of days per year in which sulfur dioxide levels exceed 150 mg/per cubic meter in Milan, Italy, is 29. The number of days per year in which emission limits are exceeded is normally distributed with a standard deviation of 4.0 days. What percentage of the years would average between 21 and 37 days of excess emissions of sulfur dioxide? What percentage of the years would exceed 37 days? What percentage of the years would exceed 41 days? In what percentage of the years would there be fewer than 25 days with excess sulfur dioxide emissions?

3.23 The Runzheimer Guide publishes a list of the most inexpensive cities in the world for the business traveler. Listed are the 10 most inexpensive cities with their respective per diem costs. Use this list to calculate the Z scores for Bordeaux, Montreal, Edmonton, and Hamilton. Treat this list as a sample.

CITY	PER DIEM ($)
Hamilton, Ontario	97
London, Ontario	109
Edmonton, Alberta	111
Jakarta, Indonesia	118
Ottawa	120
Montreal	130
Halifax, Nova Scotia	132
Winnipeg, Manitoba	133
Bordeaux, France	137
Bangkok, Thailand	137

3.3 Measures of Shape

Measures of shape are *tools that can be used to describe the shape of a distribution of data.* In this section, we examine two measures of shape—skewness and kurtosis. We also look at box and whisker plots.

Measures of shape
Tools that can be used to describe the shape of a distribution of data.

Skewness
The lack of symmetry of a distribution of values.

Skewness

A distribution of data in which the right half is a mirror image of the left half is said to be *symmetrical.* One example of a symmetrical distribution is the normal distribution, or *bell curve,* which is presented in more detail in Chapter 6.

Skewness occurs when a distribution is asymmetrical or lacks symmetry. The distribution in Figure 3.16 has no skewness because it is symmetric. Figure 3.17 shows a distribution that is skewed left, or negatively skewed, and Figure 3.18 shows a distribution that is skewed right, or positively skewed.

The skewed portion is the long, thin part of the curve. Many researchers use *skewed distribution* to mean that the data are sparse at one end of the distribution and piled up at the other end. Instructors sometimes refer to a grade distribution as *skewed,* meaning that few students scored at one end of the grading scale, and many students scored at the other end.

SKEWNESS AND THE RELATIONSHIP OF THE MEAN, MEDIAN, AND MODE The concept of skewness helps to understand the relationship of the mean, median, and mode. In a unimodal distribution (distribution with a single peak or mode) that is skewed, the mode is the apex (high point) of the curve and the median is the middle value. The mean tends to be located toward the tail of the distribution, because the mean is affected by all values, including the extreme ones. Because a bell-shaped or normal distribution has no skewness, the mean, median, and mode all are at the center of the distribution. Figure 3.19 displays the relationship of the mean, median, and mode for different types of skewness.

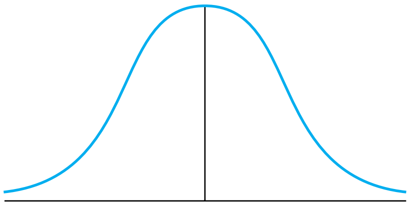


Figure 3.16

Symmetrical distribution

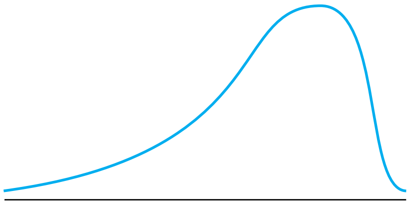


Figure 3.17

Distribution skewed left, or negatively skewed

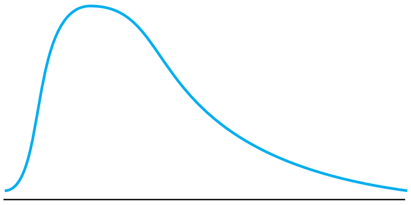


Figure 3.18

Distribution skewed right, or positively skewed

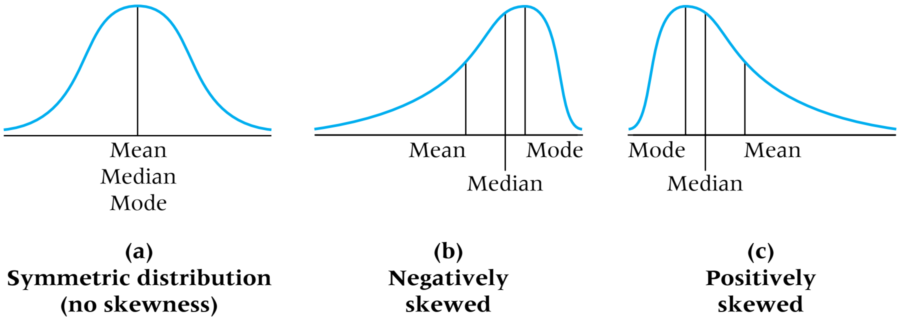


Figure 3.19

Relationship of mean, median, and mode

Coefficient of skewness
A measure of the degree of skewness that exists in a distribution of numbers; compares the mean and the median in light of the magnitude of the standard deviation.

COEFFICIENT OF SKEWNESS Statistician Karl Pearson is credited with developing at least two coefficients of skewness that can be used to determine the degree of skewness in a distribution. We present one of these coefficients here, referred to as a *Pearsonian* **coefficient of skewness.** This coefficient *compares the mean and median in light of the magnitude of the standard deviation.* Note that if the distribution is symmetrical, the mean and median are the same value and hence the coefficient of skewness is equal to zero.

COEFFICIENT OF SKEWNESS

$$S_k = \frac{3(\mu - M_d)}{\sigma}$$

where:
S_k = coefficient of skewness
M_d = median

Suppose, for example, that a distribution has a mean of 29, a median of 26, and a standard deviation of 12.3. The coefficient of skewness is computed as

$$S_k = \frac{3(29 - 26)}{12.3} = +0.73.$$

Because the value of S_k is positive, the distribution is positively skewed. If the value of S_k is negative, the distribution is negatively skewed. The greater the magnitude of S_k, the more skewed is the distribution.

Kurtosis

Kurtosis describes the *amount of peakedness of a distribution. Distributions that are high and thin* are referred to as **leptokurtic** distributions. *Distributions that are flat and spread out* are referred to as **platykurtic** distributions. Between these two types are *distributions that are more "normal" in shape,* referred to as **mesokurtic** distributions. These three types of kurtosis are illustrated in Figure 3.20.

Kurtosis
The amount of peakedness of a distribution.

Leptokurtic
Distributions that are high and thin.

Platykurtic
Distributions that are flat and spread out.

Mesokurtic
Distributions that are normal in shape—that is, not too high or too flat.

Box and whisker plot
A diagram that utilizes the upper and lower quartiles along with the median and the two most extreme values to depict a distribution graphically; sometimes called a box plot.

Box and Whisker Plots

Another way to describe a distribution of data is by using a box and whisker plot. A **box and whisker plot,** sometimes called a *box plot,* is *a diagram that utilizes the upper and lower quartiles along with the median and the two most extreme values to depict a distribution graphically.* The plot is constructed by using a box to enclose the median. This *box* is extended outward from the median along a continuum to the lower and upper quartiles, enclosing not only the median but the middle 50% of the data. From the lower and upper quartiles, lines referred to as *whiskers* are extended out from the box toward the outermost data values. The box and whisker plot is determined from five specific numbers.

1. The median (Q_2).
2. The lower quartile (Q_1).
3. The upper quartile (Q_3).
4. The smallest value in the distribution.
5. The largest value in the distribution.

The box of the plot is determined by locating the median and the lower and upper quartiles on a continuum. A box is drawn around the median with the lower and upper quartiles (Q_1 and Q_3) as the box endpoints. These box endpoints (Q_1 and Q_3) are referred to as the *hinges* of the box.

Next the value of the interquartile range (IQR) is computed by $Q_3 - Q_1$. The interquartile range includes the middle 50% of the data and should equal the length of the box. However, here the interquartile range is used outside of the box also. At a distance of $1.5 \cdot \text{IQR}$ outward from the lower and upper quartiles are what are referred to as *inner*

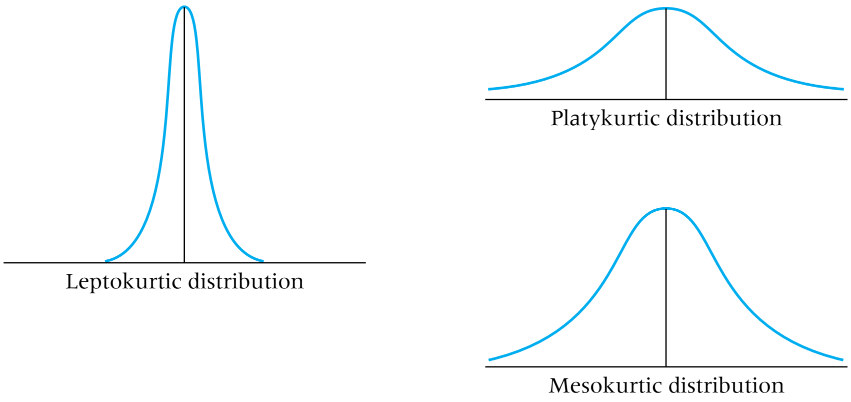


Figure 3.20

Types of kurtosis

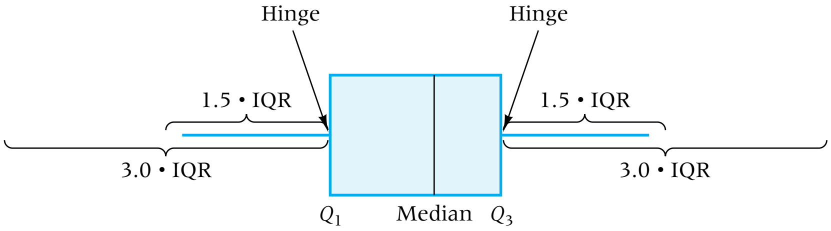


Figure 3.21

Box and whisker plot

fences. A *whisker,* a line segment, is drawn from the lower hinge of the box outward to the smallest data value. A second whisker is drawn from the upper hinge of the box outward to the largest data value. The inner fences are established as follows.

$$Q_1 - 1.5 \cdot \text{IQR}$$
$$Q_3 + 1.5 \cdot \text{IQR}$$

If there are data beyond the inner fences, then *outer fences* can be constructed:

$$Q_1 - 3.0 \cdot \text{IQR}$$
$$Q_3 + 3.0 \cdot \text{IQR}$$

Figure 3.21 shows the features of a box and whisker plot.

Data values that are outside the mainstream of values in a distribution are viewed as *outliers.* Outliers can be merely the more extreme values of a data set. However, sometimes outliers are due to measurement or recording errors. Other times they are values that are so unlike the other values that they should not be considered in the same analysis as the rest of the distribution. Values in the data distribution that are outside the inner fences but within the outer fences are referred to as *mild outliers.* Values that are outside the outer fences are called *extreme outliers.* Thus, one of the main uses of a box and whisker plot is to identify outliers.

TABLE 3.6
Data for Box and Whisker Plot

71	87	82	64	72	75	81	69
76	79	65	68	80	73	85	71
70	79	63	62	81	84	77	73
82	74	74	73	84	72	81	65
74	62	64	68	73	82	69	71

TABLE 3.7
Data in Ordered Array with Quartiles and Median

87	85	84	84	82	82	82	81	81	81
80	79	79	77	76	75	74	74	74	73
73	73	73	72	72	71	71	71	70	69
69	68	68	65	65	64	64	63	62	62

$Q_1 = 69$
$Q_2 = \text{median} = 73$
$Q_3 = 80.5$
$\text{IQR} = Q_3 - Q_1 = 80.5 - 69 = 11.5$

Another use of box and whisker plots is to determine if a distribution is skewed. If the median falls in the middle of the box, then there is no skewness. If the distribution is skewed, it will be skewed in the direction away from the median. If the median falls in the upper half of the box, then the distribution is skewed left. If the median falls in the lower half of the box, then the distribution is skewed to the right.

We shall use the data given in Table 3.6 to construct a box and whisker plot.

After organizing the data into an ordered array, as shown in Table 3.7, it is relatively easy to determine the values of the lower quartile (Q_1), the median, and the upper quartile (Q_3). From these, the value of the interquartile range can be computed.

The hinges of the box are located at the lower and upper quartiles, 69 and 80.5. The median is located within the box at distances of 4 from the lower quartile and 6.5 from the upper quartile. The distribution is skewed right, because the median is nearer to the lower or left hinge. The inner fence is constructed by

$$Q_1 - 1.5 \cdot \text{IQR} = 69 - 1.5 \cdot 11.5 = 69 - 17.25 = 51.75$$

and

$$Q_3 + 1.5 \cdot \text{IQR} = 80.5 + 1.5 \cdot 11.5 = 80.5 + 17.25 = 97.75.$$

The whiskers are constructed by drawing a line segment from the lower hinge outward to the smallest data value and a line segment from the upper hinge outward to the largest data value. An examination of the data reveals that there are no data values in this set of numbers that are outside the inner fence. The whiskers are constructed outward to the lowest value, which is 62, and to the highest value, which is 87.

To construct an outer fence, we calculate $Q_1 - 3 \cdot \text{IQR}$ and $Q_3 + 3 \cdot \text{IQR}$, as follows.

$$Q_1 - 3 \cdot \text{IQR} = 69 - 3 \cdot 11.5 = 69 - 34.5 = 34.5$$
$$Q_3 + 3 \cdot \text{IQR} = 80.5 + 3 \cdot 11.5 = 80.5 + 34.5 = 115.0$$

Analysis Using Excel

Computation of the previously presented Pearsonian coefficient of skewness is accomplished through the use of the three Excel functions, **AVERAGE, MEDIAN** and **STDEVP.**

They can be combined utilizing the Excel formula

= 3 * (AVERAGE (data range) – MEDIAN (data range)) / STDEVP (data range)

to yield the previous manually computed coefficient of skewness.

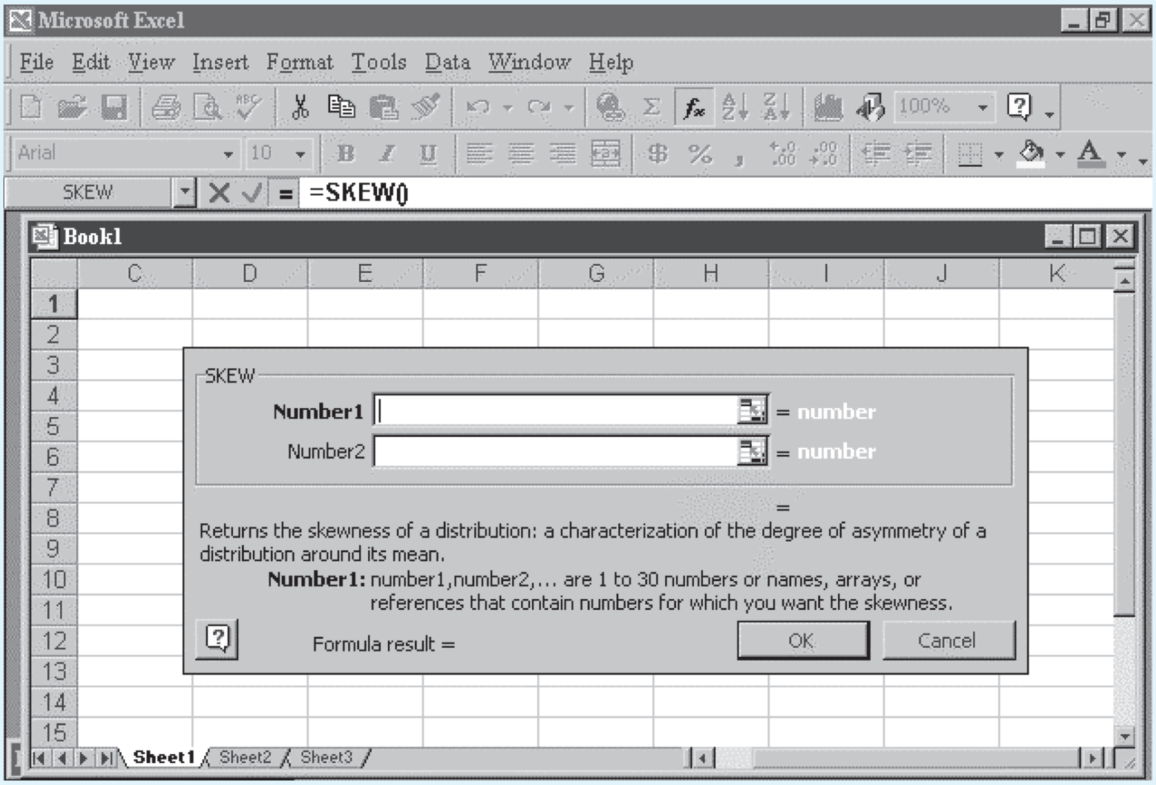
In addition, Excel has a statistical function, **SKEW,** that computes another accepted form of the coefficient of skewness. This coefficient is computed as a function of the third power of the deviations about the mean. It is accessed through the Paste Function, f_x, using **Statistical** on the left side of the dialog box and **SKEW** on the right side. Figure 3.22 displays the dialog box for **SKEW.** To use **SKEW,** insert the location of the data in the first line labeled **Number1.** The answer will appear on the dialog box; after clicking on **OK** it will appear on the spreadsheet.

Figure 3.23 displays the Excel computed value of skewness for the data from Table 3.6.

Excel cannot produce Box and Whisker Plots, but **FAST ⧹ STAT** has the capability. Figure 3.24 displays the **Box and Whisker Plot** dialog box for **FAST ⧹ STAT**. Note that the only entry requirement of this feature is the location of the data. The **FAST ⧹ STAT** Box and Whisker results consist of four general items. First, the input data values are repeated in Column A of the worksheet. Second, three output items are given. These include the box and whisker plot, the five-number summary values (smallest value, largest value, and the three quartiles) to the left of the plot, and five values on the right that are used for constructing the plot.

Shown in Figure 3.25 are two of the outputs for the data in Table 3.6. These include the five-number summary and the box and whisker plot.

Figure 3.22 Dialog box for **SKEW**



=SKEW()

SKEW

Number1 = number

Number2 = number

=

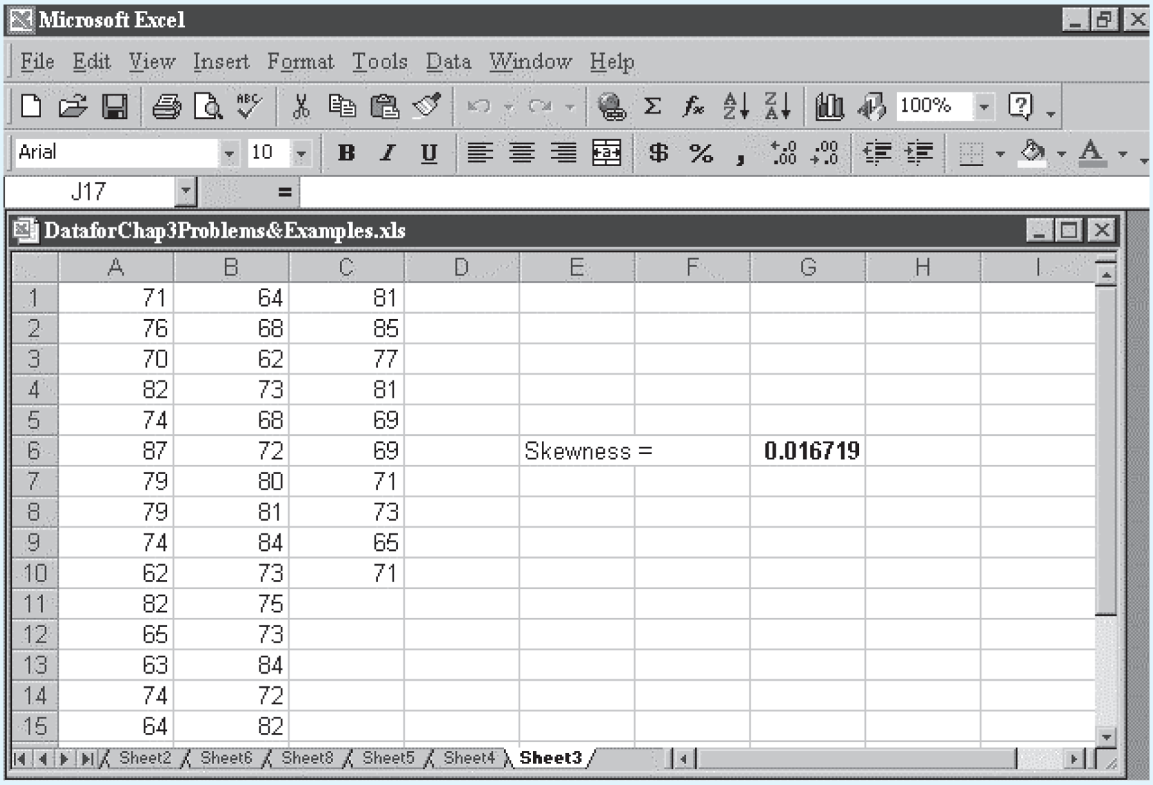
Returns the skewness of a distribution: a characterization of the degree of asymmetry of a distribution around its mean.

Number1: number1,number2,... are 1 to 30 numbers or names, arrays, or references that contain numbers for which you want the skewness.

Formula result =

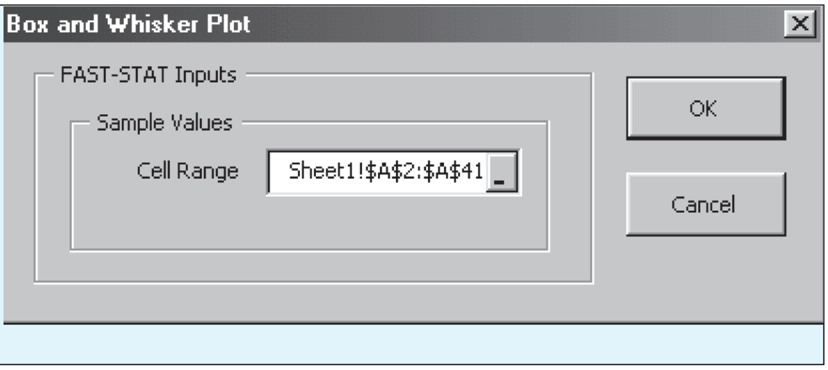
OK Cancel

Figure 3.23 Excel skewness output for the data from Table 3.6


Microsoft Excel
File Edit View Insert Format Tools Data Window Help
100%
Arial
10
J17
=
DataforChap3Problems&Examples.xls
Sheet2 Sheet6 Sheet8 Sheet5 Sheet4 Sheet3

	A	B	C	D	E	F	G	H	I
1	71	64	81						
2	76	68	85						
3	70	62	77						
4	82	73	81						
5	74	68	69						
6	87	72	69		Skewness =		0.016719		
7	79	80	71						
8	79	81	73						
9	74	84	65						
10	62	73	71						
11	82	75							
12	65	73							
13	63	84							
14	74	72							
15	64	82							

Figure 3.24 Dialog box for Box and Whisker in FAST\STAT


Box and Whisker Plot

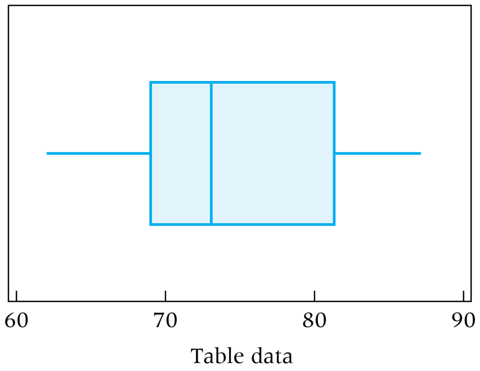
FAST-STAT Inputs

Sample Values

Cell Range: Sheet1!$A$2:$A$41

Figure 3.25 FAST\STAT box and whisker analysis of Table 3.6 data

FIVE-NUMBER SUMMARY	
Smallest Value	62
First Quartile	69
Median	73
Third Quartile	80.25
Largest Value	87



3.3 Problems

3.24 On a certain day the average closing price of a group of stocks on the New York Stock Exchange is $35 (to the nearest dollar). If the median value is $33 and the mode is $21, is the distribution of these stock prices skewed? if so, how?

3.25 A local hotel offers ballroom dancing on Friday nights. A researcher observes the customers and estimates their ages. Discuss the skewness of the distribution of ages, if the mean age is 51, the median age is 54, and the modal age is 59.

3.26 The sales volumes for the top real estate brokerage firms in the United States for a recent year were analyzed using descriptive statistics. The mean annual dollar volume for these firms was $5.51 billion, the median was $3.19 billion, and the standard deviation was $9.59 billion. Compute the value of the Pearsonian coefficient of skewness and discuss the meaning of it. Is the distribution skewed? If so, to what extent?

3.27 Suppose the data below are the ages of Internet users obtained from a sample. Use these data to compute a Pearsonian coefficient of skewness. What is the meaning of the coefficient?

41	15	31	25	24
23	21	22	22	18
30	20	19	19	16
23	27	38	34	24
19	20	29	17	23

3.28 Construct a box and whisker plot on the following data. Are there any outliers? Is the distribution of data skewed?

540	690	503	558	490	609
379	601	559	495	562	580
510	623	477	574	588	497
527	570	495	590	602	541

3.29 Suppose a consumer group asked 18 consumers to keep a yearly log of their shopping practices and that the following data represent the number of coupons used by each consumer over the yearly period. Use the data to construct a box and whisker plot. List the median, Q_1, Q_3, the endpoints for the inner fences, and the endpoints for the outer fences. Discuss the skewness of the distribution of these data and point out any outliers.

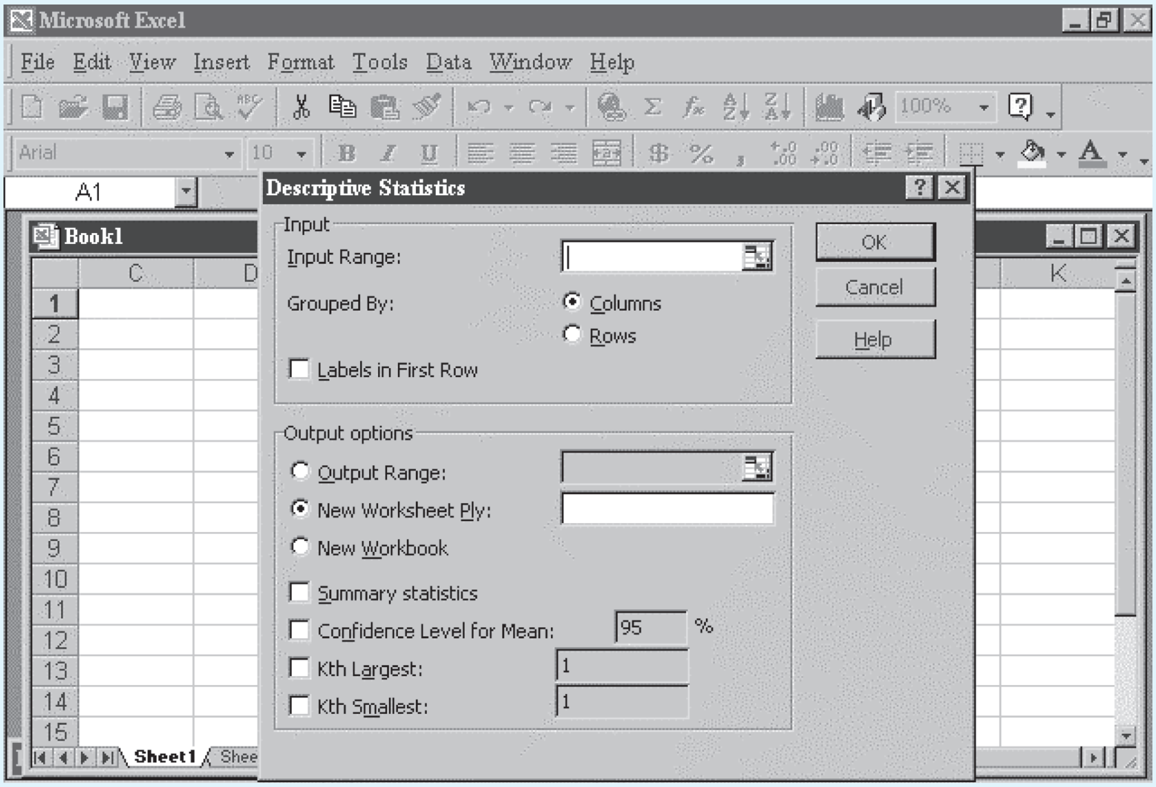
81	68	70	100	94	47	66	70	82
110	105	60	21	70	66	90	78	85

3.4 Summary Statistics in Excel

In this chapter we have introduced many descriptive statistics techniques that are useful in analyzing data. To this point we have taken an “a la carte”, one-at-a-time, Excel-approach in presenting them. However, Excel has one tool that can perform many of these functions at once. This tool is the **Descriptive Statistics** tool, which is accessed as one of the options under **_D_ata Analysis.** The **Descriptive Statistics** dialog box is displayed in Figure 3.26.

The location of the data is inserted into the **_I_nput Range** blank on the first line of the dialog box. Check the **Summary Statistics** box to calculate the several descriptive measures at once. The output contains the mean, median, mode, sample standard deviation, sample variance, range, skewness, and a measure of kurtosis. Applying this Excel feature to the data in Table 3.6 results in the output shown in Figure 3.27.

Figure 3.26 **Descriptive Statistics** dialog box



Descriptive Statistics

Input

Input Range:

Grouped By: ● Columns ○ Rows

□ Labels in First Row

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

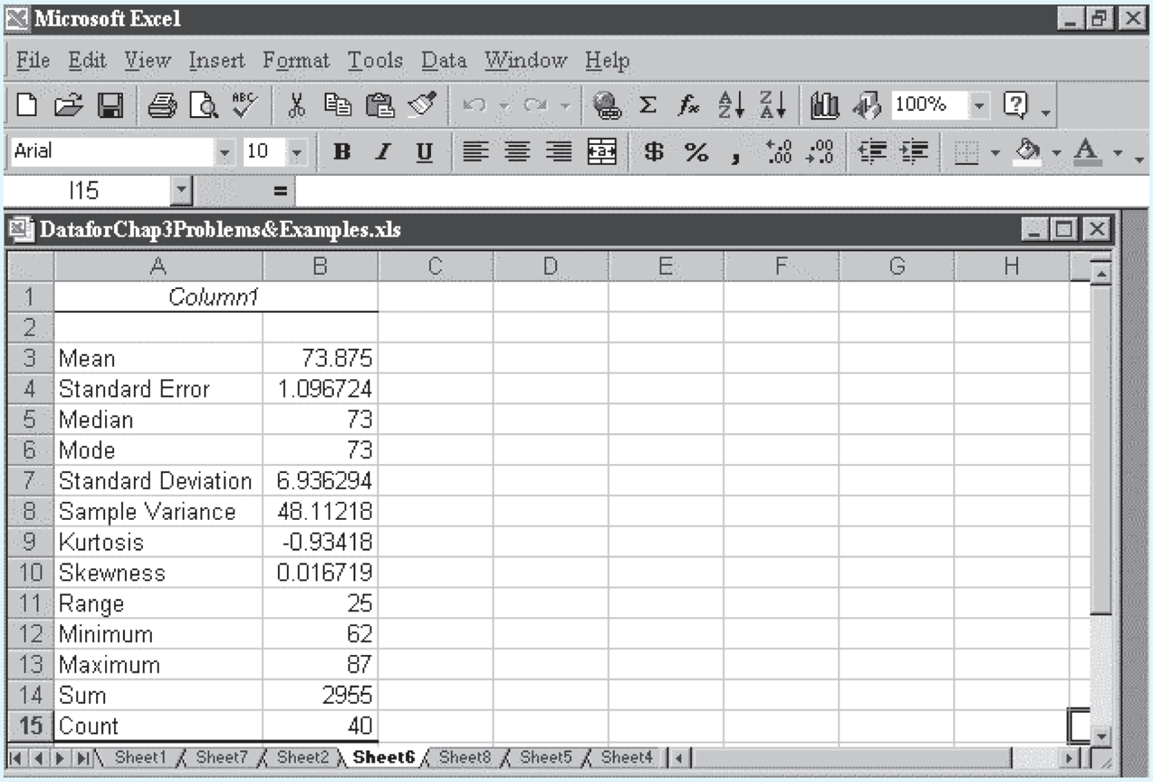
□ Summary statistics

□ Confidence Level for Mean: 95 %

□ Kth Largest: 1

□ Kth Smallest: 1

Figure 3.27 Excel **Descriptive Summary Statistics** for the Table 3.6 data



DataforChap3Problems&Examples.xls

	A	B
1	*Column1*	
2		
3	Mean	73.875
4	Standard Error	1.096724
5	Median	73
6	Mode	73
7	Standard Deviation	6.936294
8	Sample Variance	48.11218
9	Kurtosis	-0.93418
10	Skewness	0.016719
11	Range	25
12	Minimum	62
13	Maximum	87
14	Sum	2955
15	Count	40

Summary

Statistical descriptive measures include measures of central tendency, measures of variability, and measures of shape. Measures of central tendency are useful in describing data because they communicate information about the more central portions of the data. The most common measures of central tendency are the three m's: mode, median, and mean. In addition, in this text, quartiles are presented as measures of central tendency.

The mode is the most frequently occurring value in a set of data. If two values tie for the mode, the data are bimodal. Data sets can be multimodal. Among other things, the mode is used in business for determining sizes.

The median is the middle term in an ordered array of numbers if there is an odd number of terms. If there is an even number of terms, the median is the average of the two middle terms in an ordered array. The formula $(n + 1)/2$ specifies the location of the median. A median is unaffected by the magnitude of extreme values. This characteristic makes the median a most useful and appropriate measure of location in reporting such things as income, age, and prices of houses.

The arithmetic mean is widely used and is usually what researchers are referring to when they use the word *mean.* The arithmetic mean is the average. The population mean and the sample mean are computed in the same way but are denoted by different symbols. The arithmetic mean is affected by every value and can be inordinately influenced by extreme values.

Quartiles divide data into four groups. There are three quartiles: Q_1, which is the lower quartile; Q_2, which is the middle quartile and equals the median; and Q_3, which is the upper quartile.

Measures of variability are statistical tools used in combination with measures of central tendency to describe data. Measures of variability provide a description of data that measures of central tendency cannot give—information about the spread of the data values. These measures include the range, mean absolute deviation, variance, standard deviation, interquartile range, and coefficient of variation.

One of the most elementary measures of variability is the range. It is the difference between the largest and smallest values. Although the range is easy to compute, it has limited usefulness. The interquartile range is the difference between the third and first quartile. It equals the range of the middle 50% of the data.

The mean absolute deviation (MAD) is computed by averaging the absolute values of the deviations from the mean. The mean absolute deviation provides the magnitude of the average deviation but without specifying its direction. The mean absolute deviation has limited usage in statistics, but interest is growing for the use of MAD in the field of forecasting.

Variance is widely used as a tool in statistics but is little used as a stand-alone measure of variability. The variance is the average of the squared deviations about the mean.

The square root of the variance is the standard deviation. It also is a widely used tool in statistics. It is used more often than the variance as a stand-alone measure. The standard deviation is best understood by examining its applications in determining where data are in relation to the mean. The empirical rule contains statements about the proportions of data values that are within various numbers of standard deviations from the mean.

The empirical rule reveals the percentage of values that are within one, two, or three standard deviations of the mean for a set of data. The empirical rule applies only if the data are in a bell-shaped distribution. According to the empirical rule, approximately 68% of all values of a normal distribution are within plus or minus one standard deviation of the mean. Ninety-five percent of all values are within two standard deviations either side of the mean, and virtually all values are within three standard deviations of the mean. The Z score represents the number of standard deviations a value is from the mean for normally distributed data.

The coefficient of variation is a ratio of a standard deviation to its mean, given as a percentage. It is especially useful in comparing standard deviations or variances that represent data with different means.

Two measures of shape are skewness and kurtosis. Skewness is the lack of symmetry in a distribution. If a distribution is skewed, it is stretched in one direction or the other. The skewed part of a graph is its long, thin portion. One measure of skewness is the Pearsonian coefficient of skewness.

Kurtosis is the degree of peakedness of a distribution. A tall, thin distribution is referred to as leptokurtic. A flat distribution is platykurtic, and a distribution with a more normal peakedness is said to be mesokurtic.

A box and whisker plot is a graphical depiction of a distribution. The plot is constructed by using the median, the lower quartile, and the upper quartile. It can yield information about skewness and outliers.

Key Terms

arithmetic mean
bimodal
box and whisker plot
coefficient of skewness
coefficient of variation (CV)
deviation from the mean
empirical rule
interquartile range
kurtosis
leptokurtic
mean absolute deviation (MAD)
measures of central tendency
measures of shape
measures of variability
median
mesokurtic
mode
multimodal
platykurtic
quartiles
range
skewness
standard deviation
sum of squares of *X*
variance
Z score

SUPPLEMENTARY PROBLEMS

3.30 The 2000 U.S. Census asks every household to report information on each person living there. Suppose a sample of 30 households is selected and the number of persons living in each is reported as follows.

2	3	1	2	6	4	2	1	5	3	2	3	1	2	2
1	3	1	2	2	4	2	1	2	8	3	2	1	1	3

Compute the mean, median, mode, range, lower and upper quartiles, and interquartile range for these data.

3.31 The 2000 U.S. Census also asks for each person's age. Suppose that a sample of 40 households is taken from the census data and the age of the first person recorded on the census form is given as follows.

42	29	31	38	55	27	28
33	49	70	25	21	38	47
63	22	38	52	50	41	19
22	29	81	52	26	35	38
29	31	48	26	33	42	58
40	32	24	34	25		

Compute Q_1, Q_3, the interquartile range, and the range for these data.

3.32 According to the National Association of Investment Clubs, PepsiCo, Inc., is the most popular stock with investment clubs with 11,388 clubs holding PepsiCo stock. The Intel Corp. is a close second, followed by Motorola, Inc. We show a list of the most popular stocks with investment clubs. Compute the mean, median, Q_1, Q_3, range, and interquartile range for these figures.

COMPANY	NUMBER OF CLUBS HOLDING STOCK
PepsiCo, Inc.	11388
Intel Corp.	11019
Motorola, Inc.	9863
Tricon Global Restaurants	9168
Merck & Co., Inc.	8687
AFLAC Inc.	6796
Diebold, Inc.	6552

3.32 *continued*

McDonald’s Corp.	6498
Coca-Cola Co.	6101
Lucent Technologies	5563
Home Depot, Inc.	5414
Clayton Homes, Inc.	5390
RPM, Inc.	5033
Cisco Systems, Inc.	4541
General Electric Co.	4507
Johnson & Johnson	4464
Microsoft Corp.	4152
Wendy’s International, Inc.	4150
Walt Disney Co.	3999
AT&T Corp.	3619

3.33 *Editor & Publisher International Yearbook* published a listing of the top 10 daily newspapers in the United States, as shown here. Use these population data to compute a mean and a standard deviation. The figures are given in average daily circulation from Monday through Friday. Because the numbers are large, it may save you some effort to recode the data. One way to recode these data is to move the decimal point six places to the left (e.g., 1,774,880 becomes 1.77488). If you recode the data this way, the resulting mean and standard deviation will be correct for the recoded data. To rewrite the answers so that they are correct for the original data, move the decimal point back to the right six places in the answers.

NEWSPAPER	AVERAGE DAILY CIRCULATION
Wall Street Journal	1,774,880
USA Today	1,629,665
New York Times	1,074,741
Los Angeles Times	1,050,176
Washington Post	775,894
(N.Y.) *Daily News*	721,256
Chicago Tribune	653,554
Newsday	568,914
Houston Chronicle	549,101
Chicago Sun-Times	484,379

3.34 We show the companies with the largest oil refining capacity in the world according to the *Petroleum Intelligence Weekly.* Use these population data and answer the questions.

COMPANY	CAPACITY (1000s BARRELS PER DAY)
Exxon	4273
Royal Dutch/Shell	3791
China Petrochemical Corp.	2867
Petroleos de Venezuela	2437
Saudi Arabian Oil Co.	1970
British Petroleum	1965
Chevron	1661
Petrobras	1540
Texaco	1532
Petroleos Mexicanos (Pemex)	1520
National Iranian Oil Co.	1092

3.34 *continued*

a. What are the values of the mean and the median? Compare the answers and state which you prefer as a measure of location for these data and why.

b. What are the values of the range and interquartile range? How do they differ?

c. What are the values of variance and standard deviation for these data?

d. What is the *Z* score for Texaco? What is the *Z* score for Mobil? Interpret these *Z* scores.

e. Calculate the Pearsonian coefficient of skewness and comment on the skewness of this distribution.

3.35 The U.S. Department of the Interior’s Bureau of Mines releases figures on mineral production. Following are the 10 leading states in nonfuel mineral production in terms of the percentage of the U.S. total.

STATE	PERCENT OF U.S. TOTAL
Arizona	8.91
Nevada	7.69
California	7.13
Georgia	4.49
Utah	4.46
Florida	4.42
Texas	4.31
Minnesota	4.06
Michigan	3.96
Missouri	3.34

SOURCE: Bureau of Mines, U.S. Department of the Interior (*1999 World Almanac*)

a. Calculate the mean, median, and mode.

b. Calculate the range, interquartile range, mean absolute deviation, sample variance, and sample standard deviation.

c. Compute the Pearsonian coefficient of skewness for these data.

d. Sketch a box and whisker plot.

3.36 Financial analysts like to use the standard deviation as a measure of risk for a stock. The greater the deviation in a stock price over time, the more risky it is to invest in the stock. However, the average prices of some stocks are considerably higher than the average price of others, allowing for the potential of a greater standard deviation of price. For example, a standard deviation of $5.00 on a $10.00 stock is considerably different from a $5.00 standard deviation on a $40.00 stock. In this situation, a coefficient of variation might provide insight into risk. Suppose stock X costs an average of $32.00 per share and has had a standard deviation of $3.45 for the past 60 days. Suppose stock Y costs an average of $84.00 per share and has had a standard deviation of $5.40 for the past 60 days. Use the coefficient of variation to determine the variability for each stock.

3.37 The Polk Company reported that the average age of a car on U.S. roads in a recent year was 7.5 years. Suppose the distribution of ages of cars on U.S. roads is approximately bell-shaped. If 99.7% of the ages are between 1 year and 14 years, what is the standard deviation of car age? Suppose the standard deviation is 1.7 years and the mean is 7.5 years. What two values would 95% of the car ages be between?

3.38 According to a *Human Resources Report,* a worker in the industrial countries spends on average 419 minutes a day on the job. Suppose the standard deviation of time spent on the job is 27 minutes.

a. If the distribution of time spent on the job is approximately bell-shaped, between what two times would 68% of the figures be? 95%? 99.7%?

b. Suppose a worker spent 400 minutes on the job. What would that worker's Z score be and what would it tell the researcher?

3.39 During the 1990s, businesses were expected to show a lot of interest in Central and Eastern European countries. As new markets begin to open, American business people need to gain a better understanding of the market potential there. The following are the per capita GNP figures for eight of these European countries published by the *World Almanac.*

COUNTRY	PER CAPITA INCOME (U.S. $)
Albania	1290
Bulgaria	4630
Croatia	4300
Germany	20400
Hungary	7500
Poland	6400
Romania	5200
Bosnia/Herzegovina	600

a. Compute the mean and standard deviation for Albania, Bulgaria, Croatia, and Germany.

b. Compute the mean and standard deviation for Hungary, Poland, Romania, and Bosnia/Herzegovina.

c. Use a coefficient of variation to compare the two standard deviations. Treat the data as population data.

3.40 According to the Bureau of Labor Statistics, the average annual salary of a worker in Detroit, Michigan, is $35,748. Suppose the median annual salary for a worker in this group is $31,369 and the mode is $29,500. Is the distribution of salaries for this group skewed? If so, how and why? Which of these measures of central tendency would you use to describe these data? Why?

3.41 According to the U.S. Army Corps of Engineers, the top 20 U.S. ports, ranked by total tonnage (in million tons), were as follows.

PORT	TOTAL TONNAGE
Port of South Louisiana, LA	189.8
Houston, TX	148.2
New York, NY	131.6
New Orleans, LA	83.7
Baton Rouge, LA	81.0
Corpus Christi, TX	80.5
Valdez Harbor, AK	77.1
Port of Plaguemines, LA	66.9
Long Beach, CA	58.4
Texas City, TX	56.4
Mobile, AL	50.9
Pittsburgh, PA	50.9
Norfolk Harbor, VA	49.3
Tampa Harbor, FL	49.3
Lake Charles, LA	49.1
Los Angeles, CA	45.7
Baltimore Harbor, MD	43.6
Philadelphia, PA	41.9
Duluth-Superior, MN	41.4
Port Arthur, TX	37.2

a. Construct a box and whisker plot for these data.

b. Discuss the shape of the distribution from the plot.

c. Are there outliers?

d. What are they and why do you think they are outliers?

3.42 Runzheimer International publishes data on overseas business travel costs. They report that the average per diem total for a business traveler in Paris, France, is $349. Suppose the per diem costs of a business traveler to Paris are normally distributed, and 99.7% of the per diem figures are between $317 and $381. What is the value of the standard deviation? The average per diem total for a business traveler in Moscow is $415. If the shape of the distribution of per diem costs of a business traveler in Moscow is normal and if 95% of the per diem costs in Moscow lie between $371 and $459, what is the standard deviation?

ANALYZING THE DATABASES

1. Use the manufacturing database. The original data from the variable, Value of Industry Shipments, has been recoded in this database so that there are only four categories. What is the modal category? What is the mean amount of New Capital Expenditures? What is the median amount of New Capital Expenditures? What does the comparison of the mean and the median tell you about the data?
2. For the stock market database "describe" the Dollar Value variable. Include measures of central tendency, variability, and skewness. What did you find?
3. Using the financial database study Earnings per Share for Type 2 and Type 7 (chemical companies and petrochemical companies). Compute a coefficient of variability for Type 2 and for Type 7. Compare the two coefficients and comment.
4. Use the hospital database. Construct a box and whisker plot for Number of Births. Thinking about hospitals and birthing facilities, comment on why the box and whisker plot may look the way it does.

CASE

COCA-COLA GOES SMALL IN RUSSIA

The Coca-Cola Company is the number-one seller of soft drinks in the world. Every day an average of more than one billion servings of Coca-Cola, Diet Coke, Sprite, Fanta, and other products of Coca-Cola are enjoyed around the world. The company has the world's largest production and distribution system for soft drinks and sells more than twice as many soft drinks as its nearest competitor. Coca-Cola products are sold in more than 200 countries around the globe.

For several reasons, the company believes it will continue to grow internationally. One reason is that disposable income is rising. Another is that outside the United States and Europe, the world is getting younger. In addition, reaching world markets is becoming easier as political barriers fall and transportation difficulties are overcome. Still another reason is that the sharing of ideas, cultures, and news around the world creates market opportunities. Part of the company mission is for Coca-Cola to maintain the world's most powerful trademark and effectively utilize the world's most effective and pervasive distribution system.

In June 1999 Coca-Cola Russia introduced a 200-ml (about 6.8 oz) Coke bottle in Volgograd, Russia, in a campaign to market Coke to its poorest customers. This strategy was successful for Coca-Cola in other countries, such as India. The bottle sells for 12 cents, making it affordable to almost everyone.

DISCUSSION

1. Because of the variability of bottling machinery, it is likely that every bottle does not contain exactly 200 ml of fluid. Some bottles may contain more fluid and others less. Since 200-ml bottle fills are somewhat unusual, a production engineer wants to test some of the bottles from the first production runs to determine how close they are to the 200-ml specification. Suppose the following data are the fill measurements from a random sample of 50 bottles. Use the techniques presented in this chapter to describe the sample. Consider measures of central tendency, variability, and skewness. Based on this analysis, how is the bottling process working?

12.1	11.9	12.2	12.2	12.0	12.1	12.9	12.1	12.3	12.5
11.7	12.4	12.3	11.8	11.3	12.1	11.4	11.6	11.2	12.2
12.4	11.8	11.9	12.2	11.6	11.6	12.4	12.4	12.6	12.6
12.1	12.8	11.9	12.0	11.9	12.3	12.5	11.9	13.1	11.7
12.2	12.5	12.2	11.7	12.9	12.2	11.5	12.6	12.3	11.8

Suppose that at another plant Coca-Cola is filling bottles with the more traditional 20 oz of fluid. A lab randomly samples 150 bottles and tests the bottles for fill volume. The descriptive statistics are given in Excel computer output. Write a brief report to supervisors summarizing what this output is saying about the process.

Excel Output

BOTTLE FILLS

```
Mean                            20.0085
Standard error                   0.0023
Median                          20.0092
Mode                            20.0169
Standard deviation               0.0279
Sample variance                  0.0008
Kurtosis                         0.6028
Skewness                        -0.1063
Range                            0.1759
Minimum                         19.9287
Maximum                         20.1046
Sum                           3001.2707
Count                               150
```

SOURCE: Adapted from “Coke, Avis Adjust in Russia,” *Advertising Age,* July 5, 1999, p. 25, and The Coca-Cola company’s Web site at http://www.coca-cola.com/home.html.