

Technology Step by Step**Excel**
Step by Step**Finding Measures of Central Tendency****Example XL3–1**

Find the mean, mode, and median of the data from Example 3–11. The data represent the population of licensed nuclear reactors in the United States for a recent 15-year period.


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

1. On an Excel worksheet enter the numbers in cells A2–A16. Enter a label for the variable in cell A1.

On the same worksheet as the data:

2. Compute the mean of the data: key in **=AVERAGE(A2:A16)** in a blank cell.
3. Compute the mode of the data: key in **=MODE(A2:A16)** in a blank cell.
4. Compute the median of the data: key in **=MEDIAN(A2:A16)** in a blank cell.

These and other statistical functions can also be accessed without typing them into the worksheet directly.

1. Select the Formulas tab from the toolbar and select the Insert Function Icon .
2. Select the Statistical category for statistical functions.
3. Scroll to find the appropriate function and click [OK].

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

3–2**Measures of Variation**

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

Example 3–18**Objective 2**

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

**Comparison of Outdoor Paint**

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

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

Solution

The mean for brand A is

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

The mean for brand B is

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

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

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

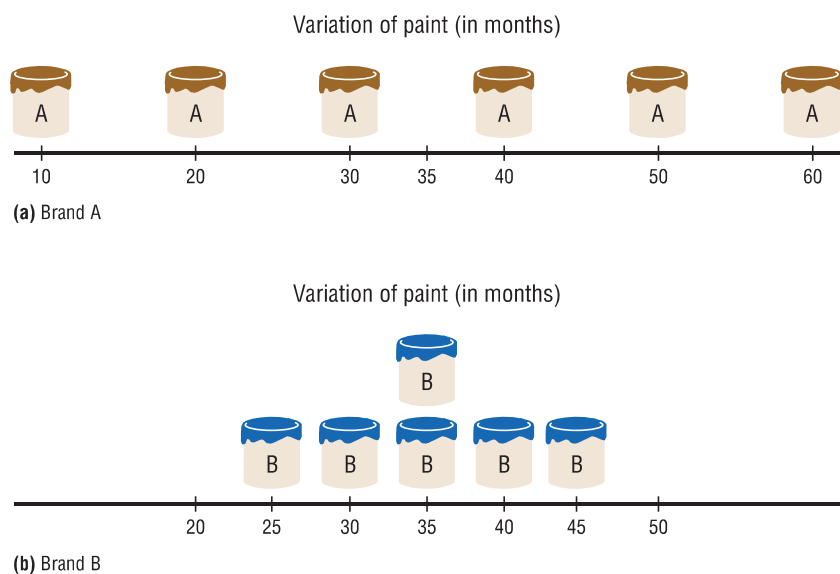
Range

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

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

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

Figure 3–2
Examining Data Sets
Graphically



Example 3–19**Comparison of Outdoor Paint**

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

Solution

For brand A, the range is

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

For brand B, the range is

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

Make sure the range is given as a single number.

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

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

Example 3–20**Employee Salaries**

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

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

Solution

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

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

Population Variance and Standard Deviation

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

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

Example 3–21**Comparison of Outdoor Paint**

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

10, 60, 50, 30, 40, 20

Solution**Step 1** Find the mean for the data.

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

Step 2 Subtract the mean from each data value.

$$\begin{array}{lll} 10 - 35 = -25 & 50 - 35 = +15 & 40 - 35 = +5 \\ 60 - 35 = +25 & 30 - 35 = -5 & 20 - 35 = -15 \end{array}$$

Step 3 Square each result.

$$\begin{array}{lll} (-25)^2 = 625 & (+15)^2 = 225 & (+5)^2 = 25 \\ (+25)^2 = 625 & (-5)^2 = 25 & (-15)^2 = 225 \end{array}$$

Step 4 Find the sum of the squares.

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

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

$$\text{Variance} = 1750 \div 6 = 291.7$$

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

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

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

Historical Note

Karl Pearson in 1892 and 1893 introduced the statistical concepts of the range and standard deviation.

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

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

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

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

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

The formula for the population variance is

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

where

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

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

The corresponding formula for the population standard deviation is

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

Example 3-22

Comparison of Outdoor Paint



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

35, 45, 30, 35, 40, 25

Solution

Step 1 Find the mean.

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

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

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

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

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

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

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

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

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

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

Hence, the standard deviation is 6.5.

Interesting Fact

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

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

Sample Variance and Standard Deviation

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

$$\frac{\sum(X - \bar{X})^2}{n}$$

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

$$\frac{\sum(X - \bar{X})^2}{n}$$

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

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

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

where

\bar{X} = sample mean

n = sample size

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

Formula for the Sample Standard Deviation

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

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

where

X = individual value

\bar{X} = sample mean

n = sample size

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

Shortcut or Computational Formulas for s^2 and s

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

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

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

Example 3–23**European Auto Sales**

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

11.2, 11.9, 12.0, 12.8, 13.4, 14.3

Source: *USA TODAY*.

Solution

Step 1 Find the sum of the values.

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

Step 2 Square each value and find the sum.

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

Step 3 Substitute in the formulas and solve.

$$\begin{aligned} s^2 &= \frac{n(\sum X^2) - (\sum X)^2}{n(n-1)} \\ &= \frac{6(958.94) - 75.6^2}{6(6-1)} \\ &= \frac{5753.64 - 5715.36}{6(5)} \\ &= \frac{38.28}{30} \\ &= 1.276 \end{aligned}$$

The variance is 1.28 rounded.

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

Hence, the sample standard deviation is 1.13.

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

Variance and Standard Deviation for Grouped Data

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

Example 3–24 **Miles Run per Week**

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

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

Solution

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

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

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

$$1 \cdot 8 = 8 \quad 2 \cdot 13 = 26 \quad \dots \quad 2 \cdot 38 = 76$$

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

$$1 \cdot 8^2 = 64 \quad 2 \cdot 13^2 = 338 \quad \dots \quad 2 \cdot 38^2 = 2888$$

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

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

Unusual Stat

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

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

$$\begin{aligned} s^2 &= \frac{n(\sum f \cdot X_m^2) - (\sum f \cdot X_m)^2}{n(n-1)} \\ &= \frac{20(13,310) - 490^2}{20(20-1)} \\ &= \frac{266,200 - 240,100}{20(19)} \\ &= \frac{26,100}{380} \\ &= 68.7 \end{aligned}$$

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

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

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

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

Procedure Table

Finding the Sample Variance and Standard Deviation for Grouped Data

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

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

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

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

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

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

$$s^2 = \frac{n(\sum f \cdot X_m^2) - (\sum f \cdot X_m)^2}{n(n-1)}$$

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

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

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

Table 3-2 Summary of Measures of Variation

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

Uses of the Variance and Standard Deviation

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

Historical Note

Karl Pearson devised the coefficient of variation to compare the deviations of two different groups such as the heights of men and women.

Coefficient of Variation

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

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

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

For samples,

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

For populations,

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

Example 3–25**Sales of Automobiles**

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

Solution

The coefficients of variation are

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

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

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

Example 3–26**Pages in Women's Fitness Magazines**

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

Solution

The coefficients of variation are

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

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

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

Range Rule of Thumb

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

The Range Rule of Thumb

A rough estimate of the standard deviation is

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

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

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

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

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

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

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

Chebyshev's Theorem

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

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

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

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

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

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

and

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

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

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

and

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

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

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

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

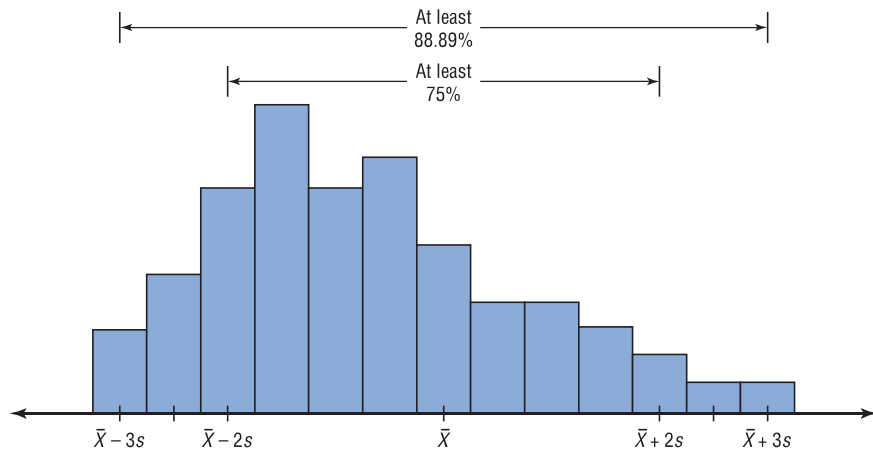
$$70 + 3(1.5) = 70 + 4.5 = 74.5$$

and

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

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

Figure 3-3
Chebyshev's Theorem



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

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

Example 3-27

Prices of Homes

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

Solution

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

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

and

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

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

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

Example 3-28

Travel Allowances

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

Solution**Step 1** Subtract the mean from the larger value.

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

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

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

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

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

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

The Empirical (Normal) Rule

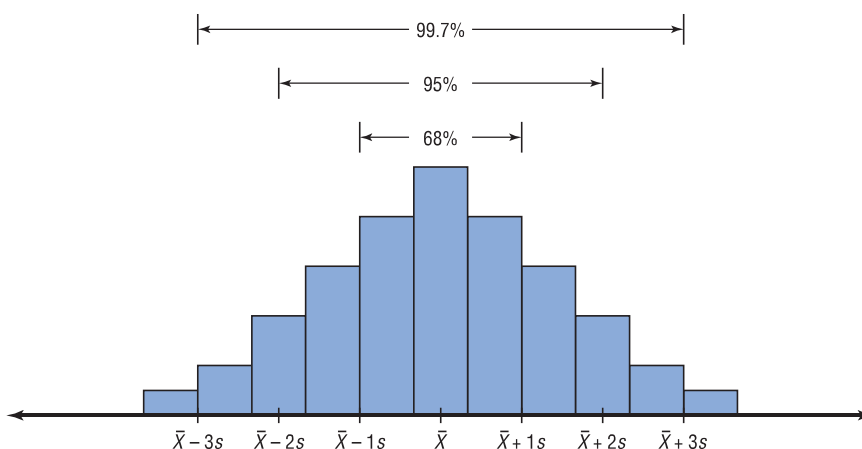
Chebyshev's theorem applies to any distribution regardless of its shape. However, when a distribution is *bell-shaped* (or what is called *normal*), the following statements, which make up the **empirical rule**, are true.

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

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

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

For example, suppose that the scores on a national achievement exam have a mean of 480 and a standard deviation of 90. If these scores are normally distributed, then approximately 68% will fall between 390 and 570 ($480 + 90 = 570$ and $480 - 90 = 390$). Approximately 95% of the scores will fall between 300 and 660 ($480 + 2 \cdot 90 = 660$ and $480 - 2 \cdot 90 = 300$). Approximately 99.7% will fall between 210 and 750 ($480 + 3 \cdot 90 = 750$ and $480 - 3 \cdot 90 = 210$). See Figure 3–4. (The empirical rule is explained in greater detail in Chapter 6.)

Figure 3–4**The Empirical Rule**

Applying the Concepts 3–2

Blood Pressure

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

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


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

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

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


See page 180 for the answers.

Exercises 3–2

1. What is the relationship between the variance and the standard deviation? *The square root of the variance is the standard deviation.*
2. Why might the range *not* be the best estimate of variability? *One extremely high or one extremely low data value will influence the range.*
3. What are the symbols used to represent the population variance and standard deviation? σ^2 ; σ
4. What are the symbols used to represent the sample variance and standard deviation? s^2 ; s
5. Why is the unbiased estimator of variance used?
6.  The three data sets have the same mean and range, but is the variation the same? Prove your answer by computing the standard deviation. Assume the data were obtained from samples.
 - a. 5, 7, 9, 11, 13, 15, 17
 - b. 5, 6, 7, 11, 15, 16, 17
 - c. 5, 5, 5, 11, 17, 17, 17 *No, a has the smallest variation; c has the biggest variation.*

For Exercises 7–17, find the range, variance, and standard deviation unless the question asks for something different. Assume the data represent samples, and use the shortcut formula for the unbiased estimator to compute the variance and standard deviation.

7. **Police Calls in Schools** The number of incidents in which police were needed for a sample of 10 schools in Allegheny County is 7, 37, 3, 8, 48, 11, 6, 0, 10, 3. Are the data consistent or do they vary? Explain your answer. *48; 254.7; 15.9 (rounded to 16) The data vary widely.*
Source: U.S. Department of Education.

8.  **Cigarette Taxes** The increases (in cents) in cigarette taxes for 17 states in a 6-month period are 60, 20, 40, 40, 45, 12, 34, 51, 30, 70, 42, 31, 69, 32, 8, 18, 50
Use the range rule of thumb to estimate the standard deviation. Compare the estimate to the actual standard deviation. *62; 332.4; 18.2; using the range rule of thumb, $s \approx 15.5$. This is close to the actual standard deviation of 18.2.*
Source: Federation of Tax Administrators.

- 9. Precipitation and High Temperatures** The normal daily high temperatures (in degrees Fahrenheit) in January for 10 selected cities are as follows.

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

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

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

Which set is more variable?

Source: *New York Times Almanac*.



- 10. Size of U.S. States** The total surface area (in square miles) for each of six selected Eastern states is listed here.

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

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

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

Which set is more variable?

Source: *New York Times Almanac*.



- 11. Stories in the Tallest Buildings** The number of stories in the 13 tallest buildings for two different cities is listed below. Which set of data is more variable?

Houston: 75, 71, 64, 56, 53, 55, 47, 55, 52, 50, 50, 50, 47

Pittsburgh: 64, 54, 40, 32, 46, 44, 42, 41, 40, 40, 34, 32, 30

Source: *World Almanac*.



- 12. Starting Teachers' Salaries** Starting teachers' salaries (in equivalent U.S. dollars) for upper secondary education in selected countries are listed below. Which set of data is more variable? (The U.S. average starting salary at this time was \$29,641.)

Europe		Asia	
Sweden	\$48,704	Korea	\$26,852
Germany	41,441	Japan	23,493
Spain	32,679	India	18,247
Finland	32,136	Malaysia	13,647
Denmark	30,384	Philippines	9,857
Netherlands	29,326	Thailand	5,862
Scotland	27,789		

Source: *World Almanac*.

- 13.** The average age of U.S. astronaut candidates in the past has been 34, but candidates have ranged in age from 26 to 46. Use the range rule of thumb to estimate the standard deviation of the applicants' ages.

Source: www.nasa.gov $s \approx R/4$ so $s \approx 5$ years.

- 14. Times Spent in Rush-Hour Traffic** A sample of 12 drivers shows the time that they spent (in minutes) stopped in rush-hour traffic on a specific snowy day last winter. *a.* 22 *b.* 35.5 *c.* 5.96

52	56	53
61	49	51
53	58	53
60	71	58

- 15. Football Playoff Statistics** The number of yards gained in NFL playoff games by rookie quarterbacks is shown. *a.* 160 *b.* 1984.5 *c.* 44.5

193	66	136	140
157	163	181	226
135	199		



- 16. Passenger Vehicle Deaths** The number of people killed in each state from passenger vehicle crashes for a specific year is shown. *a.* 2721 *b.* 355,427.6 *c.* 596.2

778	309	1110	324	705
1067	826	76	205	152
218	492	65	186	712
193	262	452	875	82
730	1185	2707	1279	390
305	123	948	343	602
69	451	951	104	985
155	450	2080	565	875
414	981	2786	82	793
214	130	396	620	797

Source: National Highway Traffic Safety Administration.

- 17.** Find the range, variance, and standard deviation for the data in Exercise 17 of Section 2–1. *a.* 46 *b.* 77.48 *c.* 8.8

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

- 18. Baseball Team Batting Averages** Team batting averages for major league baseball in 2005 are represented below. Find the variance and standard deviation for each league. Compare the results.

NL		AL	
0.252–0.256	4	0.256–0.261	2
0.257–0.261	6	0.262–0.267	5
0.262–0.266	1	0.268–0.273	4
0.267–0.271	4	0.274–0.279	2
0.272–0.276	1	0.280–0.285	1

Source: *World Almanac*. NL: $s^2 = 0.00004$, $s = 0.0066$
AL: $s^2 = 0.0000476$, $s = 0.0069$

- 19. Cost per Load of Laundry Detergents** The costs per load (in cents) of 35 laundry detergents tested by a consumer organization are shown here. 133.6; 11.6

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

- 20. Automotive Fuel Efficiency** Thirty automobiles were tested for fuel efficiency (in miles per gallon). This frequency distribution was obtained. 25.7; 5.1

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

- 21. Murders in Cities** The data show the number of murders in 25 selected cities. 27,941.46; 167.2

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

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

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

0.847; 0.920

- 23. FM Radio Stations** A random sample of 30 states shows the number of low-power FM radio stations for each state.

Class limits	Frequency
1–9	5
10–18	7
19–27	10
28–36	3
37–45	3
46–54	2

Source: Federal Communications Commission. 167.2; 12.93

- 24. Murder Rates** The data represent the murder rate per 100,000 individuals in a sample of selected cities in the United States. 134.3; 11.6

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

Source: FBI and U.S. Census Bureau.

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

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

Can it be concluded that the lifetimes of these brands of batteries are consistent? 211.2; 14.5; no, the variability of the lifetimes of the batteries is quite large.

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

- 27. Word Processor Repairs** This frequency distribution represents the data obtained from a sample of word processor repairers. The values are the days between service calls on 80 machines. 11.7; 3.4

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

- 28. Missing Work** The average number of days construction workers miss per year is 11. The standard deviation is 2.3. The average number of days factory workers miss per year is 8 with a standard deviation of 1.8. Which class is more variable in terms of days missed?

- 29. Suspension Bridges** The lengths (in feet) of the main span of the longest suspension bridges in the United States and the rest of the world are shown below. Which set of data is more variable?

United States: 4205, 4200, 3800, 3500, 3478, 2800, 2800, 2310
World: 6570, 5538, 5328, 4888, 4626, 4544, 4518, 3970

Source: *World Almanac*.

- 30. Hospital Emergency Waiting Times** The mean of the waiting times in an emergency room is 80.2 minutes with a standard deviation of 10.5 minutes for people who are admitted for additional treatment. The mean waiting time for patients who are discharged after receiving treatment is 120.6 minutes with a standard deviation of 18.3 minutes. Which times are more variable?

- 31. Ages of Accountants** The average age of the accountants at Three Rivers Corp. is 26 years, with a standard deviation of 6 years; the average salary of the accountants is \$31,000, with a standard deviation of \$4000. Compare the variations of age and income. 23.1%; 12.9%; age is more variable.

32. Using Chebyshev's theorem, solve these problems for a distribution with a mean of 80 and a standard deviation of 10.
- At least what percentage of values will fall between 60 and 100? **75%**
 - At least what percentage of values will fall between 65 and 95? **56%**
33. The mean of a distribution is 20 and the standard deviation is 2. Use Chebyshev's theorem.
- At least what percentage of the values will fall between 10 and 30? **96%**
 - At least what percentage of the values will fall between 12 and 28? **93.75%**
34. In a distribution of 160 values with a mean of 72, at least 120 fall within the interval 67–77. Approximately what percentage of values should fall in the interval 62–82? Use Chebyshev's theorem. **At least 93.75%**
35. **Calories** The average number of calories in a regular-size bagel is 240. If the standard deviation is 38 calories, find the range in which at least 75% of the data will lie. Use Chebyshev's theorem. **Between 164 and 316 calories**
36. **Time Spent Online** Americans spend an average of 3 hours per day online. If the standard deviation is 32 minutes, find the range in which at least 88.89% of the data will lie. Use Chebyshev's theorem.
Source: www.cs.cmu.edu **Between 84 and 276 minutes**
37. **Solid Waste Production** The average college student produces 640 pounds of solid waste each year. If the standard deviation is approximately 85 pounds, within what weight limits will at least 88.89% of all students' garbage lie? **Between 385 and 895 pounds**
Source: Environmental Sustainability Committee, www.esc.mtu.edu

38. **Sale Price of Homes** The average sale price of new one-family houses in the United States for 2003 was \$246,300. Find the range of values in which at least 75% of the sale prices will lie if the standard deviation is \$48,500. **Between \$149,300 and \$343,300**
Source: *New York Times Almanac*.

39. **Trials to Learn a Maze** The average of the number of trials it took a sample of mice to learn to traverse a maze was 12. The standard deviation was 3. Using Chebyshev's theorem, find the minimum percentage of data values that will fall in the range of 4–20 trials. **86%**

40. **Farm Sizes** The average farm in the United States in 2004 contained 443 acres. The standard deviation is 42 acres. Use Chebyshev's theorem to find the minimum percentage of data values that will fall in the range of 338–548 acres. **At least 84%**
Source: *World Almanac*.


41. **Citrus Fruit Consumption** The average U.S. yearly per capita consumption of citrus fruit is 26.8 pounds. Suppose that the distribution of fruit amounts consumed is bell-shaped with a standard deviation equal to 4.2 pounds. What percentage of Americans would you expect to consume more than 31 pounds of citrus fruit per year? **16%**
Source: USDA/Economic Research Service.

42. **Work Hours for College Faculty** The average full-time faculty member in a post-secondary degree-granting institution works an average of 53 hours per week.

- If we assume the standard deviation is 2.8 hours, what percentage of faculty members work more than 58.6 hours a week? **No more than 12.5%**
- If we assume a bell-shaped distribution, what percentage of faculty members work more than 58.6 hours a week? **2.5%**


Source: National Center for Education Statistics.

Extending the Concepts

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

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

All the data values fall within 2 standard deviations of the mean.

-  44. **Ages of Consumers** For this data set, find the mean and standard deviation of the variable. The data represent the ages of 30 customers who ordered a product advertised on television. Count the number of data values that fall within 2 standard deviations of the mean. Compare this with the number obtained from Chebyshev's theorem. Comment on the answer. **93.3%; All but two data values fall within 2 standard deviations of the mean.**

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

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

k	1.5	2	2.5	3	3.5
Percent	56	75	84	88.89	92

46. Use this data set: 10, 20, 30, 40, 50
- Find the standard deviation. 15.81
 - Add 5 to each value, and then find the standard deviation. 15.81
 - Subtract 5 from each value and find the standard deviation. 15.81
 - Multiply each value by 5 and find the standard deviation. 79.06
 - Divide each value by 5 and find the standard deviation. 3.16
 - Generalize the results of parts b through e .
 - Compare these results with those in Exercise 38 of Exercises 3-1.



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

$$\text{Mean deviation} = \frac{\sum |X - \bar{X}|}{n}$$

where

X = value

\bar{X} = mean

n = number of values

$|$ = absolute value

Find the mean deviation for these data.

5, 9, 10, 11, 11, 12, 15, 18, 20, 22 4.36

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

$$PC = \frac{3(\bar{X} - MD)}{s}$$

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

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

- Mean = 10, median = 8, standard deviation = 3.
 - Mean = 42, median = 45, standard deviation = 4.
 - Mean = 18.6, median = 18.6, standard deviation = 1.5.
 - Mean = 98, median = 97.6, standard deviation = 4.
49. All values of a data set must be within $s\sqrt{n-1}$ of the mean. If a person collected 25 data values that had a mean of 50 and a standard deviation of 3 and you saw that one data value was 67, what would you conclude?

Technology Step by Step

Excel Step by Step

Finding Measures of Variation


Example XL3-2

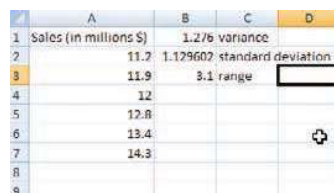
Find the variance, standard deviation, and range of the data from Example 3-23. The data represent the amount (in millions of dollars) of European auto sales for a sample of 6 years.

11.2 11.9 12.0 12.8 13.4 14.3

- On an Excel worksheet enter the data in cells A2:A7. Enter a label for the variable in cell A1.
- For the sample variance, enter =VAR(A2:A7).
- For the sample standard deviation, enter =STDEV(A2:A7).
- For the range, compute the difference between the maximum and the minimum values by entering =MAX(A2:A7) - MIN(A2:A7).

These and other statistical functions can also be accessed without typing them into the worksheet directly.

- Select the Formulas tab from the toolbar and select the Insert Function Icon .
- Select the Statistical category for statistical functions.
- Scroll to find the appropriate function and click [OK].



	A	B	C	D
1	Sales (in millions \$)	1.275 variance		
2	11.2	1.129607 standard deviation		
3	11.9	3.1 range		
4	12			
5	12.8			
6	13.4			
7	14.3			
8				
9				