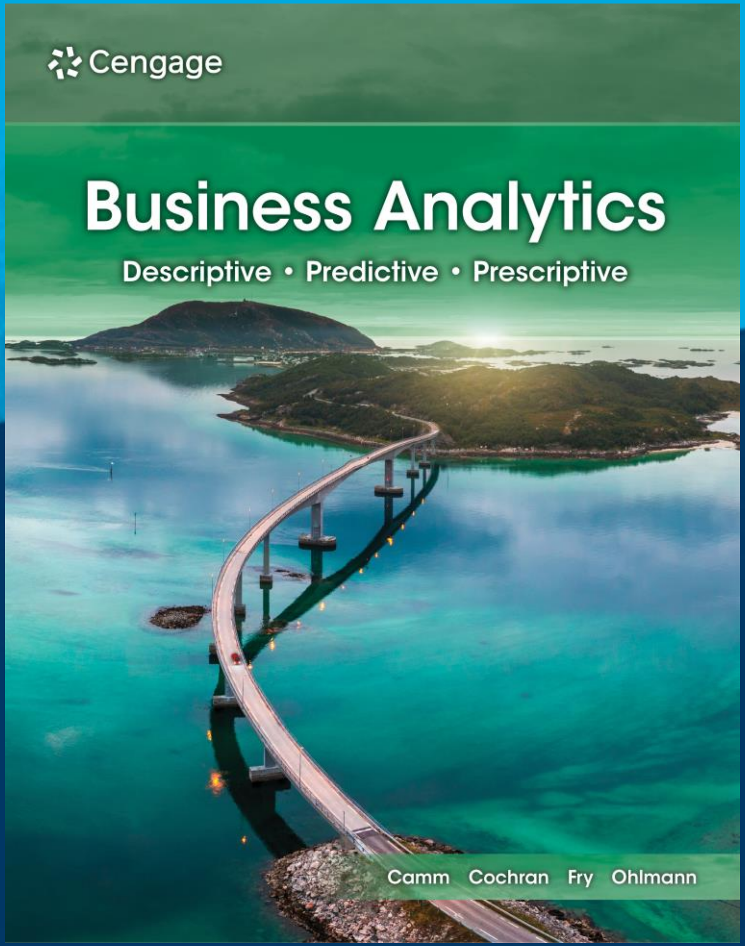


Business Analytics, 5e

Chapter 2 – Descriptive Statistics

CENGAGE

Chapter Contents

2.1 Overview of Using Data: Definitions and Goals

2.2 Types of Data

2.3 Exploring Data in Excel

2.4 Creating Distributions from Data

2.5 Measures of Location

2.6 Measures of Variability

2.7 Analyzing Distributions

2.8 Measures of Association Between Two Variables

Summary

CENGAGE

Learning Objectives (1 of 3)

After completing this chapter, you will be able to:

LO 2-1 Identify and describe different data types, including population and sample data, quantitative and categorical data, and cross-sectional and time-series data.

LO 2-2 Generate insights through sorting, filtering, and conditional formatting data.

LO 2-3 Construct and interpret frequency, relative frequency, and percent frequency distributions for categorical data.

LO 2-4 Construct and interpret frequency, relative frequency, and percent frequency distributions for quantitative data.

LO 2-5 Construct and interpret histograms and frequency polygons to visualize the distribution of quantitative data.

Learning Objectives (2 of 3)

LO 2-6 Construct and interpret cumulative frequency, cumulative relative frequency, and cumulative percent frequency distributions for quantitative data.

LO 2-7 Interpret the shape of a distribution of data and identify positive skewness, negative skewness, and symmetric distributions.

LO 2-8 Calculate and interpret measures of location such as the mean, median, geometric mean, and mode.

LO 2-9 Calculate and interpret measures of variability such as the range, variance, standard deviation, and coefficient of variation.

LO 2-10 Analyze and interpret distributions of data using percentiles, quartiles, z-scores, and the empirical rule.

Learning Objectives (3 of 3)

LO 2-11 Identify outliers in a set of data.

LO 2-12 Construct and interpret a boxplot.

LO 2-13 Create and interpret a scatter chart for two quantitative variables.

LO 2-14 Calculate and interpret the covariance and correlation coefficient for two quantitative variables.

2.1 Overview of Using Data: Definitions and Goals

Data are the facts and figures collected, analyzed, and summarized for presentation and interpretation.

A **variable** is a characteristic or a quantity of interest that can take on different values.

An **observation** is a set of values corresponding to a set of variables.

Variation is the difference in a variable measured over observations.

The role of descriptive analytics is to collect and analyze data to better understand variation and its impact on the business setting.

A **random variable**, or **uncertain variable**, is a quantity whose values are not known with certainty.

2.2 Types of Data

A **sample** is a subset of the **population**, which consists of all the elements of interest.

- **Random sampling** is a sampling method that allows gathering a representative sample from the population data.

Data are considered **quantitative data** if numeric and arithmetic operations, such as addition, subtraction, multiplication, and division, can be performed.

If arithmetic operations cannot be performed, they are **categorical data**.

Cross-sectional data are data collected from several entities at the same, or approximately the same, point in time.

Time series data are data collected over several time periods.

2.2 Sources of Data

Data necessary to analyze a business problem can often be obtained with a statistical study.

- Statistical studies can be classified as experimental or observational.

In an **experimental study,** a variable of interest is first identified.

- Then, one or more other variables are identified and controlled or manipulated to obtain data about how they influence the variable of interest.

A **nonexperimental** or **observational study** does not attempt to control the variables of interest.

- A survey is perhaps the most common type of observational study.

2.3 Sorting Data in Excel

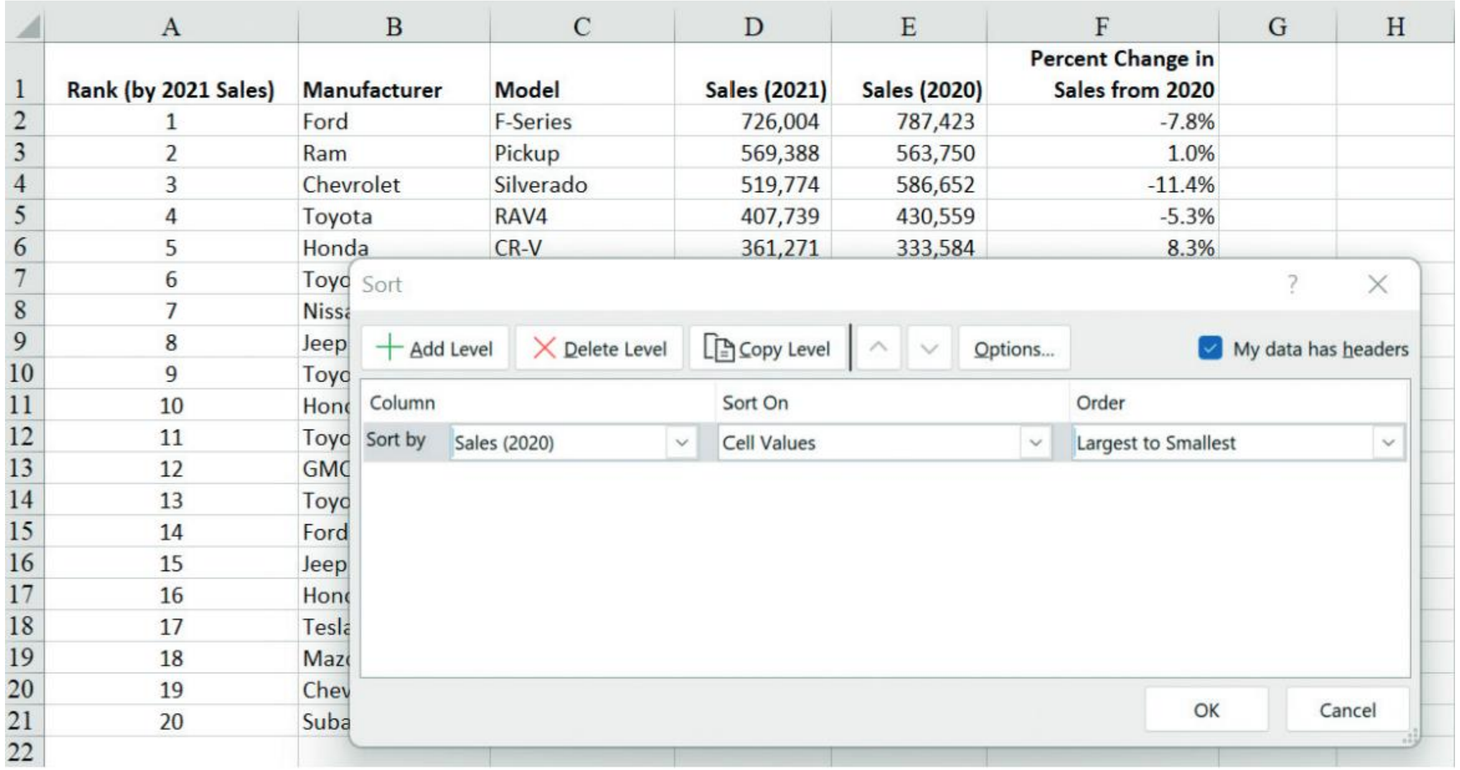
DATAfile *top20vehicles2021* contains 2020 and 2021 sales for the 20 top-selling vehicles in the United States. The data are sorted by 2021 sales.

Column F contains the percentage of sales from 2020.

Cell F2 is calculated as:

=(D2−E2)/D2

To sort the data by 2020 sales, follow the steps included in the notes.



	A	B	C	D	E	F	G	H
1	Rank (by 2021 Sales)	Manufacturer	Model	Sales (2021)	Sales (2020)	Percent Change in Sales from 2020		
2	1	Ford	F-Series	726,004	787,423	-7.8%		
3	2	Ram	Pickup	569,388	563,750	1.0%		
4	3	Chevrolet	Silverado	519,774	586,652	-11.4%		
5	4	Toyota	RAV4	407,739	430,559	-5.3%		
6	5	Honda	CR-V	361,271	333,584	8.3%		
7	6	Toyo						
8	7	Niss						
9	8	Jeep						
10	9	Toyo						
11	10	Hon						
12	11	Toyo						
13	12	GMC						
14	13	Toyo						
15	14	Ford						
16	15	Jeep						
17	16	Hon						
18	17	Tesl						
19	18	Maz						
20	19	Chev						
21	20	Suba						
22								

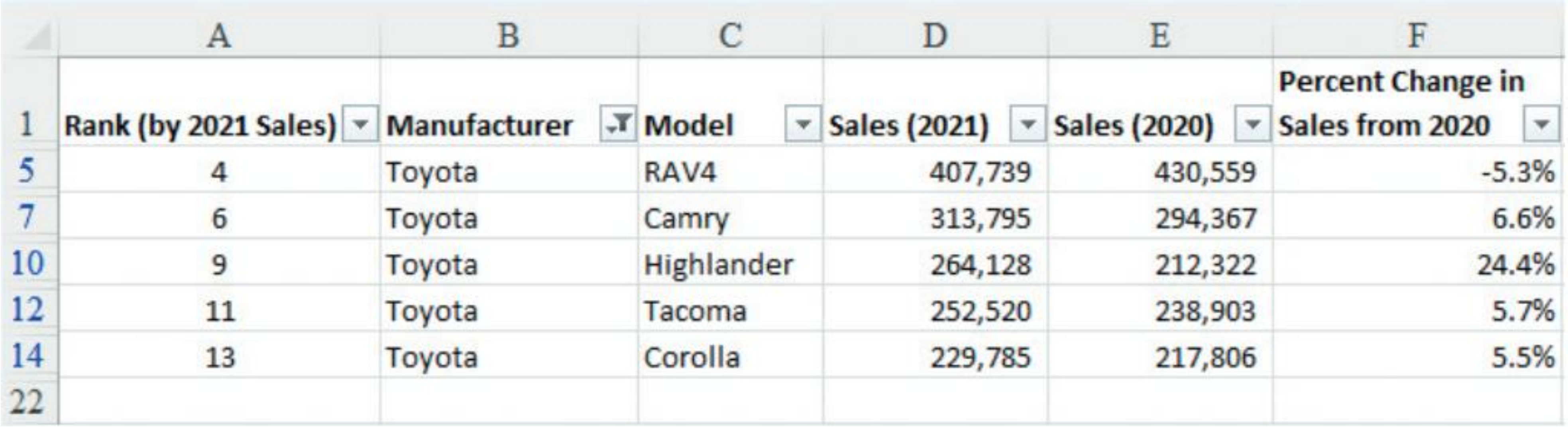
☑ My data has headers

	Column	Sort On	Order
Sort by	Sales (2020)	Cell Values	Largest to Smallest

2.3 Filtering Data in Excel

Now let’s suppose that we are interested only in seeing the sales of models made by Toyota.

To view only Toyota models, use Excel’s Filter function as shown in the notes.



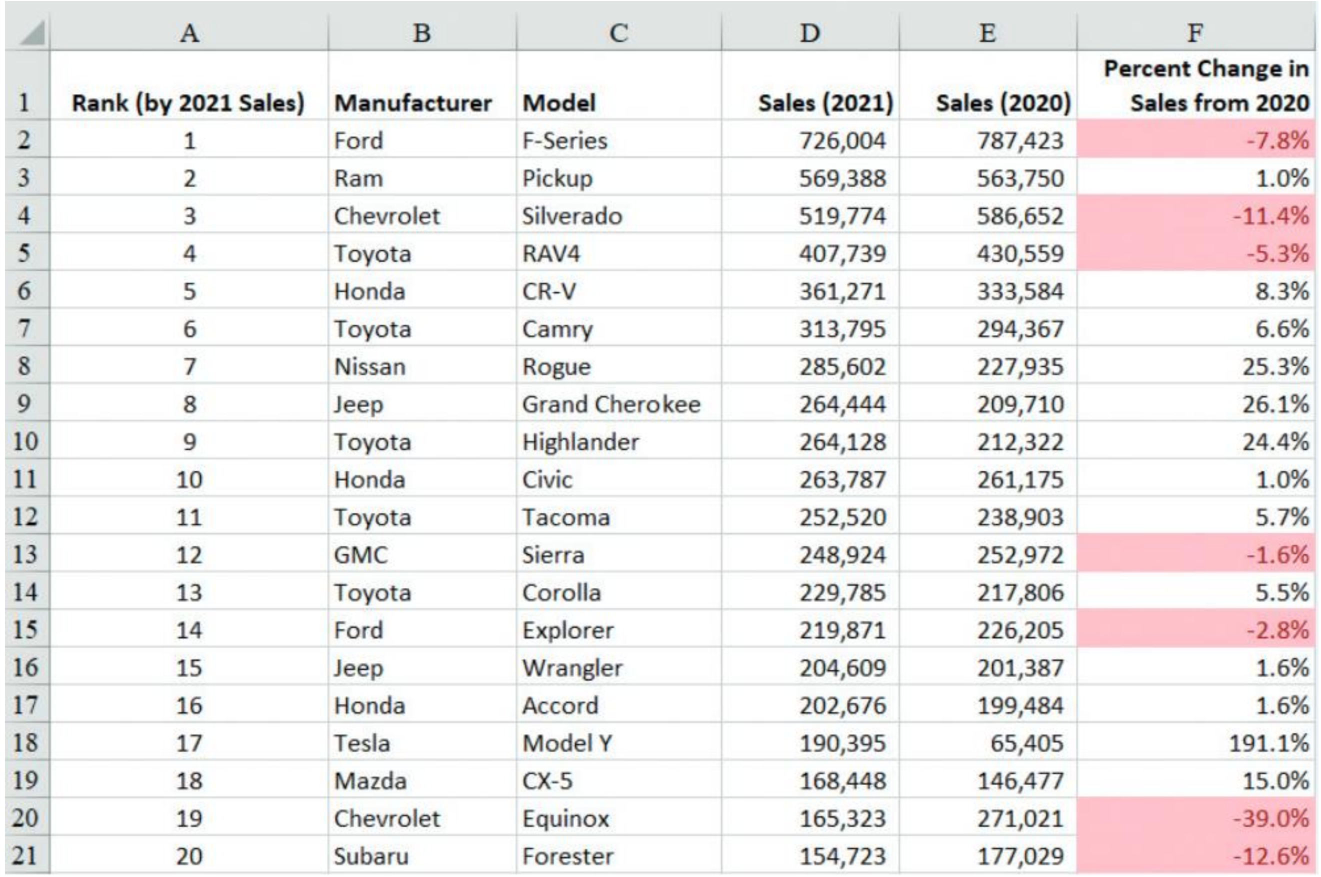
	A	B	C	D	E	F
1	Rank (by 2021 Sales)	Manufacturer	Model	Sales (2021)	Sales (2020)	Percent Change in Sales from 2020
5	4	Toyota	RAV4	407,739	430,559	-5.3%
7	6	Toyota	Camry	313,795	294,367	6.6%
10	9	Toyota	Highlander	264,128	212,322	24.4%
12	11	Toyota	Tacoma	252,520	238,903	5.7%
14	13	Toyota	Corolla	229,785	217,806	5.5%
22						

2.3 Conditional Formatting in Excel

Conditional formatting makes it easy to identify data that satisfy certain conditions in a data set.

To identify the models for which sales decreased from 2020 to 2021, follow the steps shown in the notes.

See the text for an example of using conditional formatting to generate data bars for the top-selling vehicles.



	A	B	C	D	E	F
1	Rank (by 2021 Sales)	Manufacturer	Model	Sales (2021)	Sales (2020)	Percent Change in Sales from 2020
2	1	Ford	F-Series	726,004	787,423	-7.8%
3	2	Ram	Pickup	569,388	563,750	1.0%
4	3	Chevrolet	Silverado	519,774	586,652	-11.4%
5	4	Toyota	RAV4	407,739	430,559	-5.3%
6	5	Honda	CR-V	361,271	333,584	8.3%
7	6	Toyota	Camry	313,795	294,367	6.6%
8	7	Nissan	Rogue	285,602	227,935	25.3%
9	8	Jeep	Grand Cherokee	264,444	209,710	26.1%
10	9	Toyota	Highlander	264,128	212,322	24.4%
11	10	Honda	Civic	263,787	261,175	1.0%
12	11	Toyota	Tacoma	252,520	238,903	5.7%
13	12	GMC	Sierra	248,924	252,972	-1.6%
14	13	Toyota	Corolla	229,785	217,806	5.5%
15	14	Ford	Explorer	219,871	226,205	-2.8%
16	15	Jeep	Wrangler	204,609	201,387	1.6%
17	16	Honda	Accord	202,676	199,484	1.6%
18	17	Tesla	Model Y	190,395	65,405	191.1%
19	18	Mazda	CX-5	168,448	146,477	15.0%
20	19	Chevrolet	Equinox	165,323	271,021	-39.0%
21	20	Subaru	Forester	154,723	177,029	-12.6%

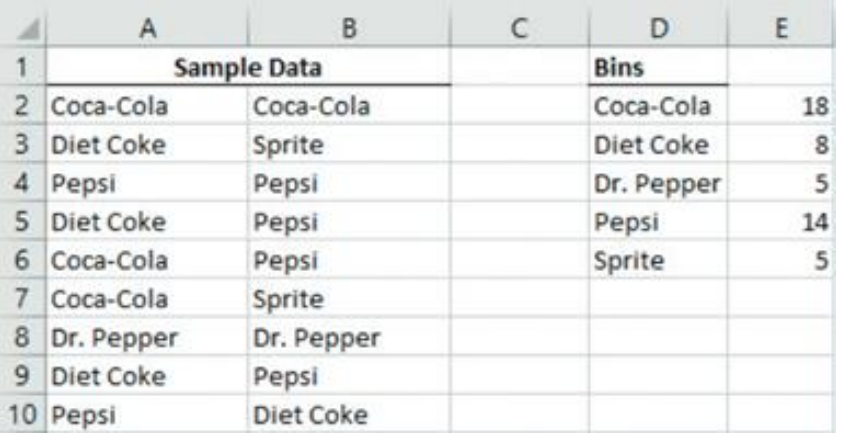
2.4 Frequency Distributions for Categorical Data

A **frequency distribution** is a summary of data showing the number (frequency) of observations in several nonoverlapping classes, typically referred to as **bins**.

DATAfile: *softdrinks*

We obtain the frequency distribution of soft drink purchases shown in cell range D2:E6 by counting the number of times each soft drink appears in the data (follow the steps included in the notes.)

The frequency distribution in cell range D2:E6 summarizes how the 50 soft drink purchases are distributed across the five soft drinks.



	A	B	C	D	E
1	Sample Data			Bins	
2	Coca-Cola	Coca-Cola		Coca-Cola	18
3	Diet Coke	Sprite		Diet Coke	8
4	Pepsi	Pepsi		Dr. Pepper	5
5	Diet Coke	Pepsi		Pepsi	14
6	Coca-Cola	Pepsi		Sprite	5
7	Coca-Cola	Sprite			
8	Dr. Pepper	Dr. Pepper			
9	Diet Coke	Pepsi			
10	Pepsi	Diet Coke			

Soft Drink	Frequency
Coca-Cola	18
Diet Coke	8
Dr. Pepper	5
Pepsi	14
Sprite	5
Total	50

2.4 Relative Frequency and Percent Frequency Distributions

For a data set with *n* observations, the **relative frequency** of each bin can be determined as follows:

$$\textbf{Relative Frequency of a bin} = \frac{\textbf{Frequency of the bin}}{\boldsymbol{n}}$$

A **relative frequency distribution** is a tabular summary of data showing the relative frequency for each bin.

A **percent frequency distribution** summarizes the percent frequency of the data for each bin.

- It can be used to provide estimates of the relative likelihoods of different values of a random variable.

Soft Drink	**Relative Frequency**	**Percent Frequency (%)**
Coca-Cola	0.36	36
Diet Coke	0.16	16
Dr. Pepper	0.10	10
Pepsi	0.28	28
Sprite	0.10	10
Total	1.00	100

2.4 Frequency Distributions for Quantitative Data

The three steps necessary to define the classes for a frequency distribution with quantitative data are as follows:

1. Determine the number of nonoverlapping bins.
   - We recommend using 5 to 20 bins.
2. Determine the width of each bin.

$$\text{Approximate bin width} = \frac{\text{Largest Data Value} - \text{Smallest data value}}{\text{Number of bins}}$$

3. Determine the range spanned by the set of bins.
   - The lower and upper bin limits must be chosen so that each data item belongs to one and only one class.

CENGAGE

2.4 Frequency Distribution for the *Age of Death* Data

Number of bins: because of the large data set (n=700), we select 16 bins.

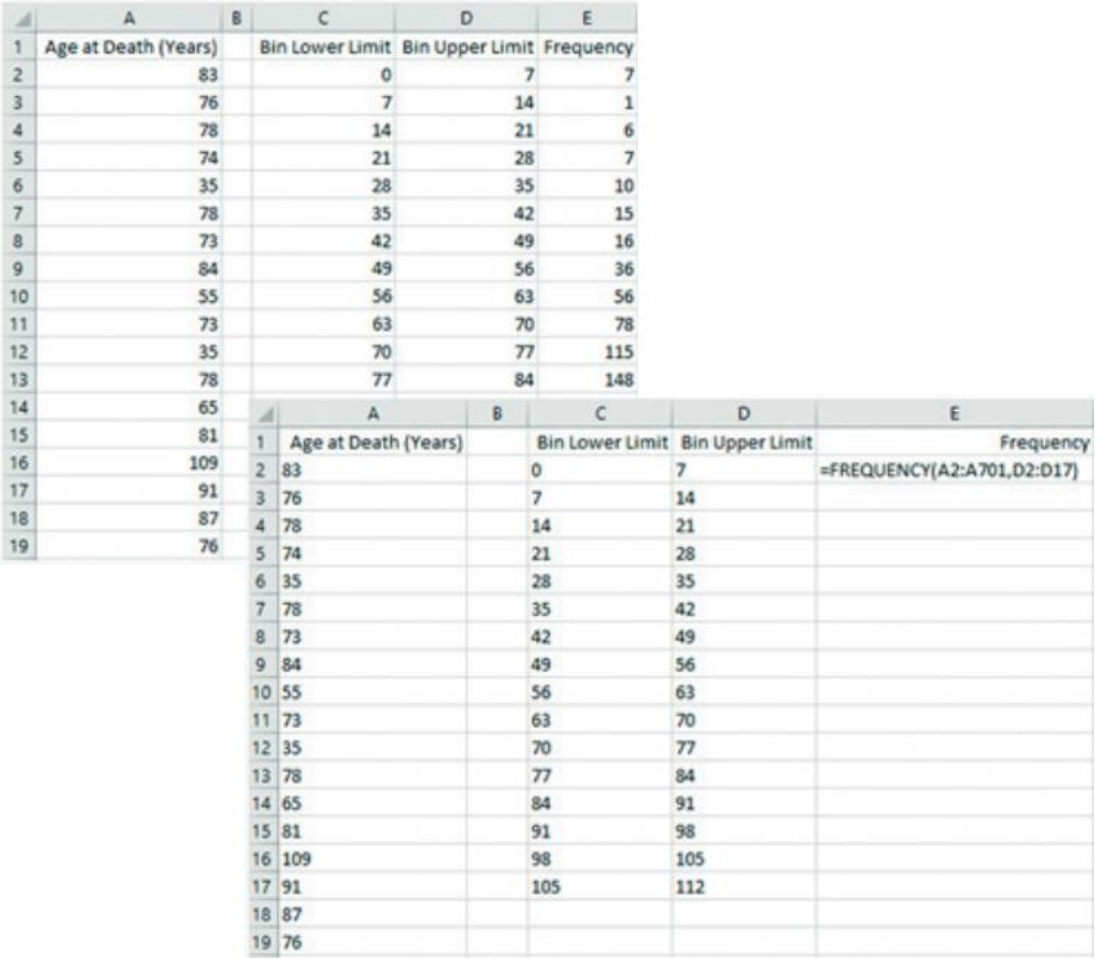
Width of bins $= (109 - 0)/16 \approx 7$

With $6 \times 17 = 112$, we choose the *bin range* [0-112], with nonoverlapping bin limits: [0,7], (7,14], (14,21]… (105,112].

The data are in cell range A1:A701.

Columns C and D contain the lower and upper bin limits.

The steps to calculate the frequencies in column E are shown in the notes.



	A	B	C	D	E
1	Age at Death (Years)		Bin Lower Limit	Bin Upper Limit	Frequency
2	83		0	7	7
3	76		7	14	1
4	78		14	21	6
5	74		21	28	7
6	35		28	35	10
7	78		35	42	15
8	73		42	49	16
9	84		49	56	36
10	55		56	63	56
11	73		63	70	78
12	35		70	77	115
13	78		77	84	148
14	65				
15	81				
16	109				
17	91				
18	87				
19	76				

	A	B	C	D	E
1	Age at Death (Years)		Bin Lower Limit	Bin Upper Limit	Frequency
2	83		0	7	=FREQUENCY(A2:A701,D2:D17)
3	76		7	14	
4	78		14	21	
5	74		21	28	
6	35		28	35	
7	78		35	42	
8	73		42	49	
9	84		49	56	
10	55		56	63	
11	73		63	70	
12	35		70	77	
13	78		77	84	
14	65		84	91	
15	81		91	98	
16	109		98	105	
17	91		105	112	
18	87				
19	76				

2.4 Histograms

A **Histogram** is a common graphical presentation of quantitative data.

- A histogram is constructed by placing the variable of interest on the horizontal axis and the selected frequency measure (absolute frequency, relative frequency, or percent frequency) on the vertical axis.
- The frequency measure of each class is shown by drawing a rectangle whose base is the class limits on the horizontal axis and whose height is the corresponding frequency measure.
- A histogram is a column chart with no spaces between the columns whose heights represent the frequencies of the corresponding bins.
  - Eliminating the space between the columns allows a histogram to reflect the continuous nature of the variable of interest.

CENGAGE

2.4 Histogram for the *Age of Death* Data

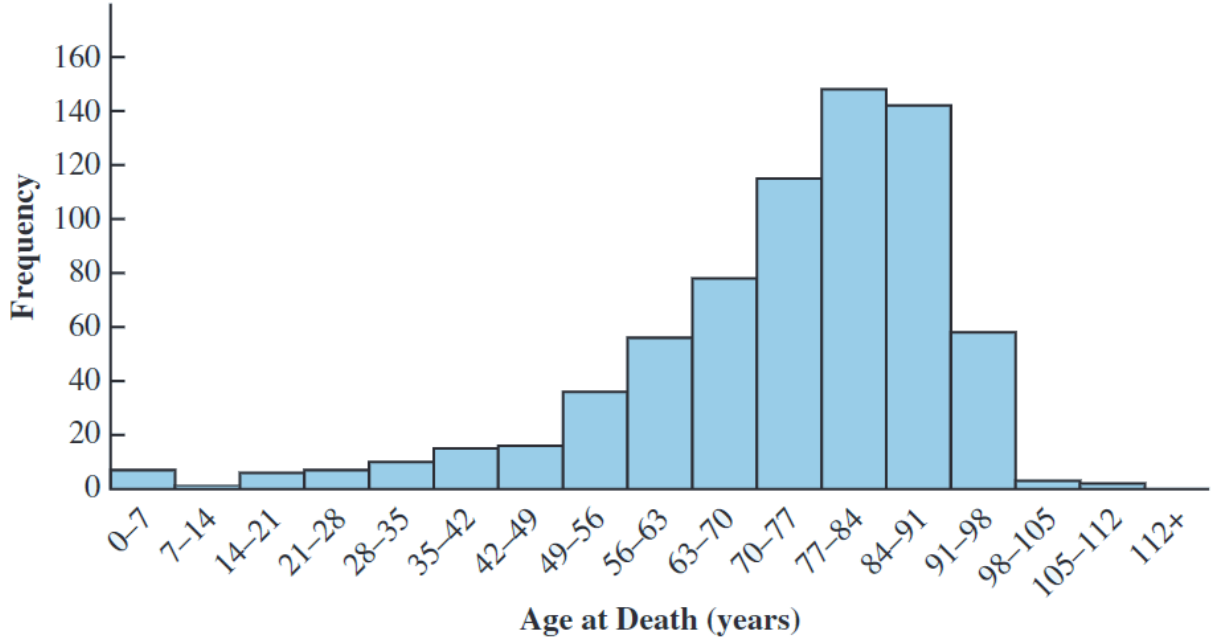
A histogram can be automatically generated using Excel’s Charts functionality.

- Select cell range A1-A701, and click the **Insert Statistics Chart** button located in the **Charts** group of the **Insert** Tab.

Excel automatically chooses to use 16 bins, each spanning 7 years, traversing the range from 0 to 112.

The notes show how to generate a histogram manually from a column chart.

Use of the function CONCAT allows to generate the horizontal labels.



2.4 Frequency Polygons

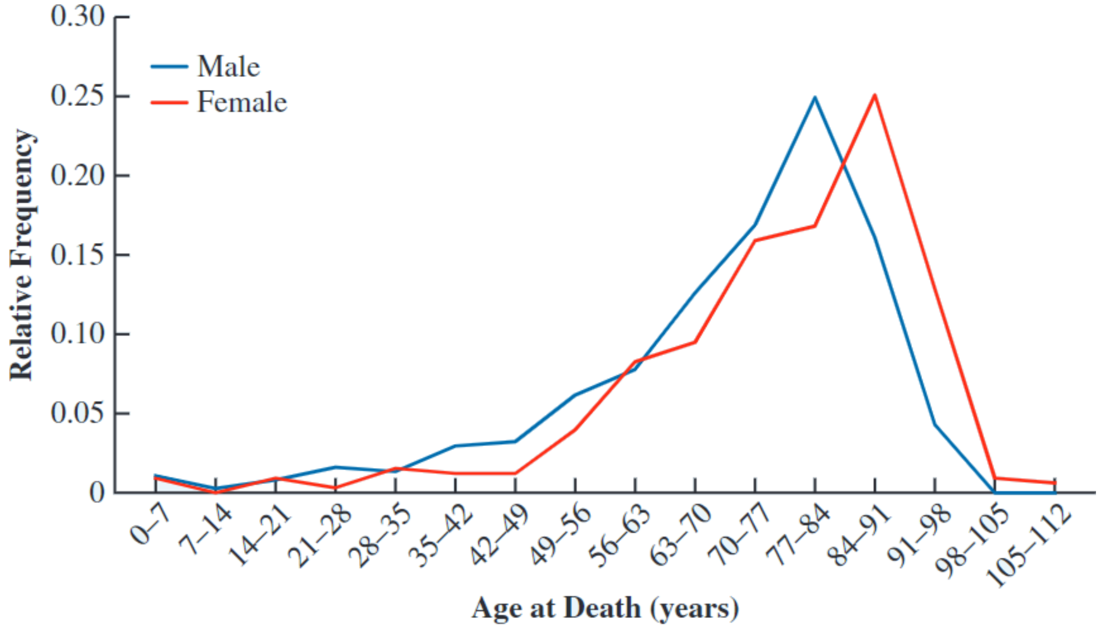
A **frequency polygon** is useful for comparing quantitative distributions.

- A frequency polygon uses lines to connect the frequency counts of observations from different bins.

DATAfile: *agedeath_sex*

The data set contains the same ages for 700 individuals we saw before, but the top 327 observations are from individuals who self-identify as females and the bottom 373 as males.

The notes show how to generate a frequency polygon in Excel.



2.4 Cumulative Distributions

A **cumulative frequency distribution** is a variation of the frequency distribution that provides another tabular summary of quantitative data.

- It uses the number of classes, class widths, and class limits developed for the frequency distribution.
- Shows the number of data items with values less than or equal to the upper-class limit of each class.

The table to the right shows the cumulative distributions for the age of death data.

Age at Death (Years)	Cumulative Frequency	Cumulative Relative Frequency	Cumulative Percent Frequency (%)
≤7	7	0.010	1.0
≤14	8	0.011	1.1
≤21	14	0.020	2.0
≤28	21	0.030	3.0
≤35	31	0.044	4.4
≤42	46	0.066	6.6
≤49	62	0.089	8.9
≤56	98	0.140	14.0
≤63	154	0.220	22.0
≤70	232	0.331	33.1
≤77	347	0.496	49.6
≤84	495	0.707	70.7
≤91	637	0.910	91.0
≤98	695	0.993	99.3
≤105	698	0.997	99.7
≤112	700	1.000	100.0

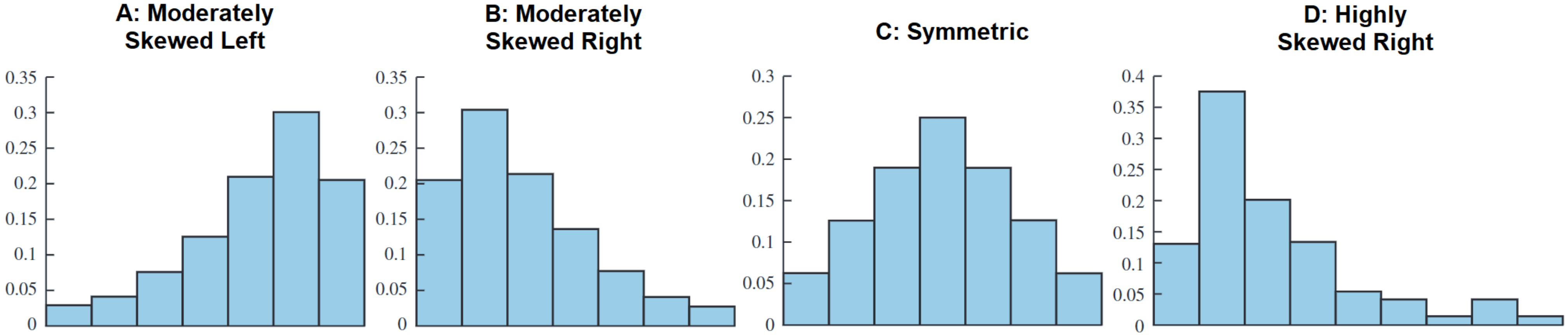
CENGAGE

2.4 Histograms Showing Differing Levels of Skewness

Histograms provide information about the shape, or form, of a distribution.

Skewness, or lack of symmetry, is an important characteristic of the shape of a distribution.

The four histograms shown below are constructed from relative frequency distributions that exhibit different patterns of skewness.



A: Moderately Skewed Left

B: Moderately Skewed Right

C: Symmetric

D: Highly Skewed Right

2.5 Mean

The most common measure of central location is the **mean**, the average of all the data values. The population mean is denoted by the Greek letter, μ.

For a sample with *n* observations, mean is computed as follows.

$$\boldsymbol{\bar{x}} = \frac{\sum \boldsymbol{x_i}}{\boldsymbol{n}} \qquad \text{where } x_i \text{ is the } i\text{th observation}$$

DATAfile: *homesales*

The mean home selling price for the sample of 12 home sales is:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{\sum x_1 + \cdots + x_{12}}{n} = \frac{138{,}000 + \cdots + 456{,}250}{12} = 219{,}937.50$$

In Excel, the value for the mean in is calculated using *=AVERAGE(B2:B13)*.

2.5 Median

The **median** is the value in the middle of a data set when data are arranged in ascending order.

To compute the median, arrange the data in ascending order. Then

a. if *n* is *odd*, the median is the middle value

b. if *n* is *even*, the median is the average of the two middle values

For the home sales data, $n = 12$ is even, and the median is computed as the average of the 6th and 7th (middle) values.

$$Median = (199{,}500 + 208{,}000)/2 = 203{,}750$$

Because extremely small and large data values influence the mean, the median is the preferred measure of central location for highly skewed data.

2.5 Mode

The **mode** of a data set is the value that occurs with the greatest frequency.

In Excel, we can find the mode using the *MODE.SNGL* function.

The greatest frequency may occur at two or more different values. In these instances, more than one mode exists.

- If the data have exactly two modes, the data are said to be *bimodal*.
- If the data have more than two modes, the data are said to be *multimodal*.

In Excel, we can find multiple modes using the *MODE.MULT* function.

- For the home sales data, *=MODE.MULT(B2:B13)* returns two modes.

2.5 Geometric Mean

The **geometric mean** is a measure of central location calculated by finding the nth root of the product of n values.

The general formula for the geometric mean, denoted $\bar{x}_g$, follows.

$$\boldsymbol{\bar{x}_g = \sqrt[n]{(x_1)(x_2)\ldots(x_n)} = [(x_1)(x_2)\ldots(x_n)]^{1/n}}$$

The geometric mean is often used in analyzing growth rates in financial data (where using the arithmetic mean will provide misleading results.)

It should be applied any time you want to determine the mean rate of change over several successive periods (be it years, quarters, weeks, etc.)

Other common applications include changes in populations of species, crop yields, pollution levels, and birth and death rates.

2.5 An Application of the Geometric Mean

DATAfile: *mutualfundreturns*

Year	Return (%)	Growth Factor
1	-22.1	0.779
2	28.7	1.287
3	10.9	1.109
4	4.9	1.049
5	15.8	1.158
6	5.5	1.055
7	-37.0	0.630
8	26.5	1.265
9	15.1	1.151
10	2.1	1.021

With a percentage annual return for year 1 of −22.1%, the balance in the fund at the end of year 1 is

$$\$100(1-0.221)=\$100(0.779)=\$77.90$$

We refer to 0.779 as the **growth factor** for year 1.

Generalizing the results, at the end of year 10, the initial investment would be worth

$$\$100[(0.779)(1.287)\ldots(1.021)]=\$100(1.3345)=\$133.45$$

Thus, the fund average annual return is (see notes for the $\bar{x}_g$ Excel formula)

$$(\bar{x}_g-1)100\%=\left(\sqrt[`10]{1.3345}-1\right)100\%=(1.0293-1)100\%=2.93\%$$

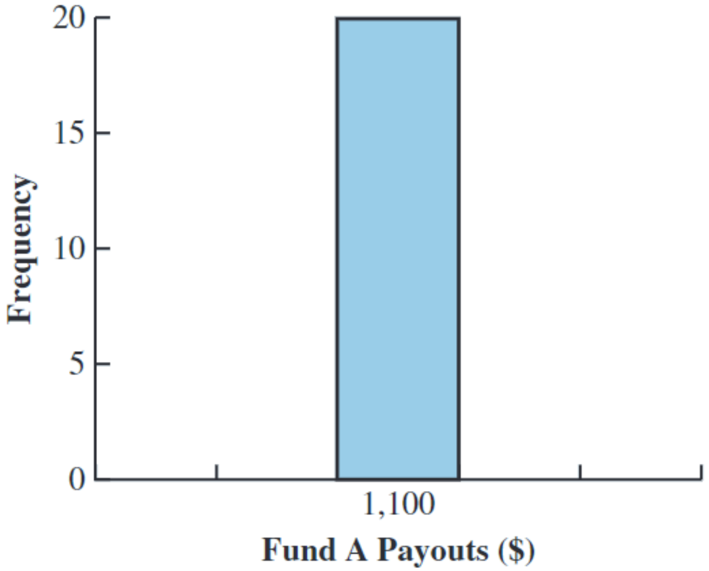
2.6 Measures of Variability

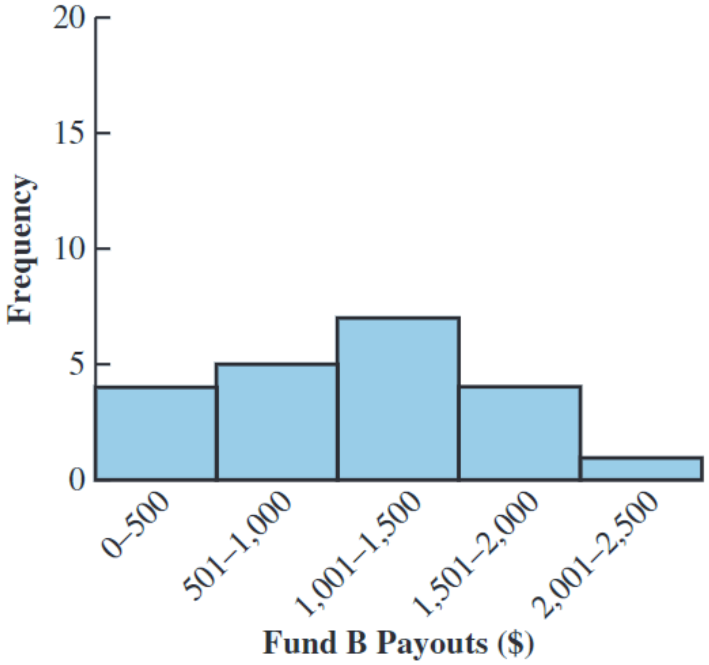
It is often desirable to consider measures of variability, or dispersion).

Consider the annual payouts for two different investment funds, A and B.

Although the mean payout is the same for the two funds, their histograms differ because the payouts associated with Fund B have greater variability.

In this section, we present several ways to measure variability.





2.6 Range

The **range** is the simplest measure of variability, and it is defined as

Range = Largest Value – Smallest Value

For the home sales data, the range is

$$\$456{,}250 - \$108{,}000 = \$348{,}250$$

In Excel, the range is computed using the *MAX* and *MIN* functions.

=MAX(B2,B13)−MIN(B2,B13)

However, the range sensitivity to extreme data values makes it a poor choice to measure the dispersion in a data set.

2.6 Variance

The **variance** is a measure of variability that utilizes all the data.

The variance is based on the *deviation about the mean*, written as $(x_i - \bar{x})$.

In most statistical applications, when we compute a sample variance, we are often interested in using it to estimate the unknown population variance.

For a random sample, if the sum of the squared deviations about the sample mean is divided by $n - 1$, and not n, the resulting sample variance provides an unbiased estimate of the population variance.

For this reason, the sample variance, denoted by s^2, is defined as follows.

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

CENGAGE

2.6 Computation of the Variance

Consider the data on the class size from five college classes:

46 54 42 46 32

The table shows the computations of the squared deviations about the mean, $(x_i - \bar{x})^2$.

The sample variance is computed as:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}\frac{256}{4} = 64\ (\text{students})^2$$

x_i	$\bar{x}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
46	44	2	4
54	44	10	100
42	44	-2	4
46	44	2	4
32	44	-12	144
		0	256

In Excel, the sample variance is computed using the formula *VAR.S*. For the home sales data, we have =*VAR.S(B2,B13)* = 9,037,501,420.

2.6 Standard Deviation

The positive square root of the variance is the **standard deviation**.

The sample standard deviation, s, is a point estimate of the population standard deviation, σ, and is derived from the sample variance as follows:

$$\boldsymbol{s = \sqrt{s^2}}$$

Because of the square root, the variance, $s^2 = 64\ (\text{students})^2$ in our example, is converted to $s = \sqrt{64} = 8\ students$ in the standard deviation.

- The standard deviation always has the same units as the original data.

In Excel, the sample standard deviation is computed using the formula *STDEV.S*. For the home sales data, we have =*STDEV.S(B2,B13)* = $95,065.77*.*

2.6 Coefficient of Variation

The **coefficient of variation**, usually expressed as a percentage, measures how large the standard deviation is relative to the mean.

$$\left(\frac{\textbf{Standard Deviation}}{\textbf{Mean}} \times \mathbf{100}\right)\%$$

For the class size example, $\bar{x} = 44$ and $s = 8$ students. The coefficient of variation is $(8/44 \times 100)\% = 18.2\%$

In words, the coefficient of variation tells us that the sample standard deviation is 18.2% of the value of the sample mean.

For the home sales data example, the coefficient of variation is

$$(95{,}065.77/219{,}937.50 \times 100)\% = 43.22\%$$

CENGAGE

2.7 Percentiles

The **pth percentile** of a data set is a value such that

- at least *p%* of the items take on this value or less, and
- at least $(100 – p)$% of the items take on this value or more.

To calculate the pth percentile of a data set, we must first sort the data in ascending order.

The **location** of the pth percentile, denoted L_p, is computed using the following:

$$\boldsymbol{L_p = \frac{p}{100}(n+1)}$$

Once we obtain the location, we are ready to calculate the pth percentile.

As an example, let us calculate the 85th percentile for the home sales data.

2.7 An Application of the Percentile

With the home sales data arranged in ascending order, we indicate the position of each observation directly below its value.

Price	108,000	138,000	138,000	142,000	186,000	199,500	208,000	254,000	254,000	257,500	298,000	456,250
Position	1	2	3	4	5	6	7	8	9	10	11	12

The location of the 85th percentile is

$$L_{85} = \frac{p}{100}(n+1) = \frac{85}{100}(12+1) = 11.05$$

The interpretation of $L_{85} = 11.05$ is that the 85th percentile is 5% of the way between the values in position 11 and 12 (see notes for Excel calculations.)

$$\text{85th percentile} = 298{,}000 + 0.05(456{,}250 - 298{,}000) = 305{,}912.50$$

CENGAGE

2.7 Quartiles

Quartiles are specific percentiles that divide the data set into four parts, with each part containing approximately 25% of the observations.

Quartiles are defined as follows:

$Q_1 =$ first quartile, or 25th percentile

$Q_2 =$ second quartile, or 50th percentile (also the median)

$Q_3 =$ third quartile, or 75th percentile

The difference between the third and first quartiles is often referred to as the **interquartile range**, or IQR.

The procedure for computing percentiles is also used to compute quartiles.

2.7 An Application of Quartiles

Let us calculate Q_1 and Q_3 for the home sales sample.

First, we calculate the locations:

$$L_{25} = \frac{25}{100}(12+1) = 3.25 \quad \text{and} \quad L_{75} = \frac{75}{100}(12+1) = 9.75$$

The calculations of Q_1 and Q_3 follow (see notes for the Excel function.)

$$Q_1 = 25th \text{ percentile} = 138{,}000 + 0.25(142{,}000 - 138{,}000) = 139{,}000$$

$$Q_3 = 75th \text{ percentile} = 254{,}000 + 0.75(257{,}500 - 254{,}000) = 256{,}625$$

And the interquartile range is:

$$IQR = Q_3 - Q_1 = 256{,}625 - 139{,}000 = 117{,}625$$

CENGAGE

2.7 *z*-Scores

The ***z*-score** helps us measure the relative location of a value in the data set.

A *z*-score is often called a *standardized value*, denotes the number of standard deviations, s a data value x_i is from the mean, $\bar{x}$.

$$\boldsymbol{z_i = \frac{x_i - \bar{x}}{s}}$$

The *z*-scores for the class size data are computed to the right (see notes for Excel.)

For example, for $x_1 = 46$, the *z*-score is

$$z_1 = \frac{x_1 - \bar{x}}{s} = \frac{46 - 44}{8} = \frac{2}{8} = 0.25$$

$\boldsymbol{x_i}$	$\boldsymbol{x_i - \bar{x}}$	$\boldsymbol{z_i = \frac{x_i - \bar{x}}{s}}$
46	2	2/8 = 0.25
54	10	10/8 = 1.25
42	2	−2/8 = −0.25
46	2	2/8 = 0.25
32	12	−12/8 = −1.50

2.7 Empirical Rule

When the data are believed to approximate a symmetric bell-shaped distribution, the **empirical rule** can be used to determine the percentage of data values within a specified number of standard deviations of the mean.

For data having a bell-shaped distribution:

- ~68% of the data values will be within one standard deviation of the mean.
- ~95% of the data values will be within two standard deviations of the mean.
- Almost all data values will be within three standard deviations of the mean.

If the height of adult males in the U.S. follows a bell-shaped distribution, with a mean of 69.5 inches and a standard deviation of 3 inches:

- ~95% of U.S. adult males will have heights between 63.5 and 75.5 inches.

2.7 Identifying Outliers

An **outlier** is an unusually small or unusually large value in a data set.

Care should be taken when handling outliers, as they might be:

- an incorrectly recorded data value
- a data value that was incorrectly included in the data set
- a correctly recorded data value that belongs in the data set

Standardized values (*z*-scores) can be used to identify potential outliers.

- A data value with a *z*-score less than –3 or greater than +3 might be a candidate for being an outlier (see notes for details.)
- Such data values can then be reviewed to determine their accuracy and whether they belong in the data set.

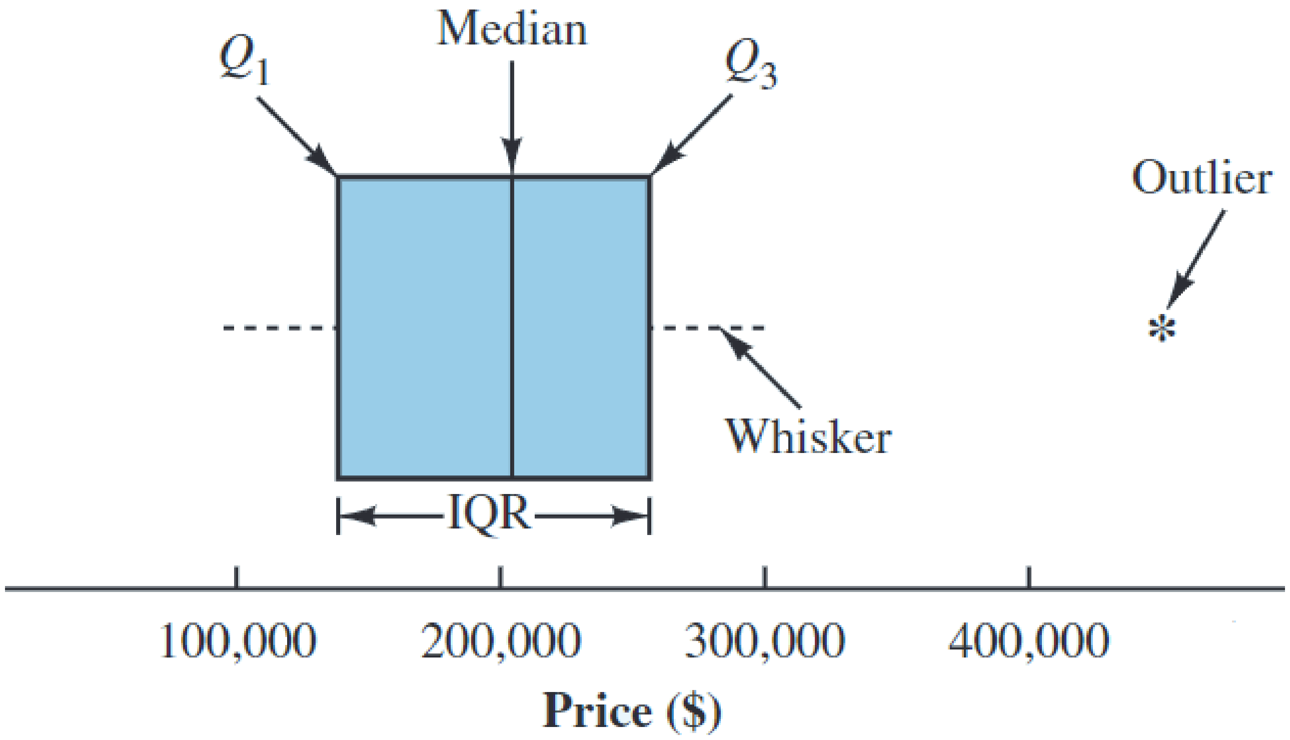
CENGAGE

2.7 Boxplots

A **boxplot**, also known as box-and-whisker plot, is a graphical summary of the distribution of data developed from the quartiles for a data set.

Shown to the right is the boxplot for the home sales data containing a single outlier.

See notes for the steps needed to create a boxplot in Excel.



2.7 Boxplots Comparing Home Sales Prices in Different Communities

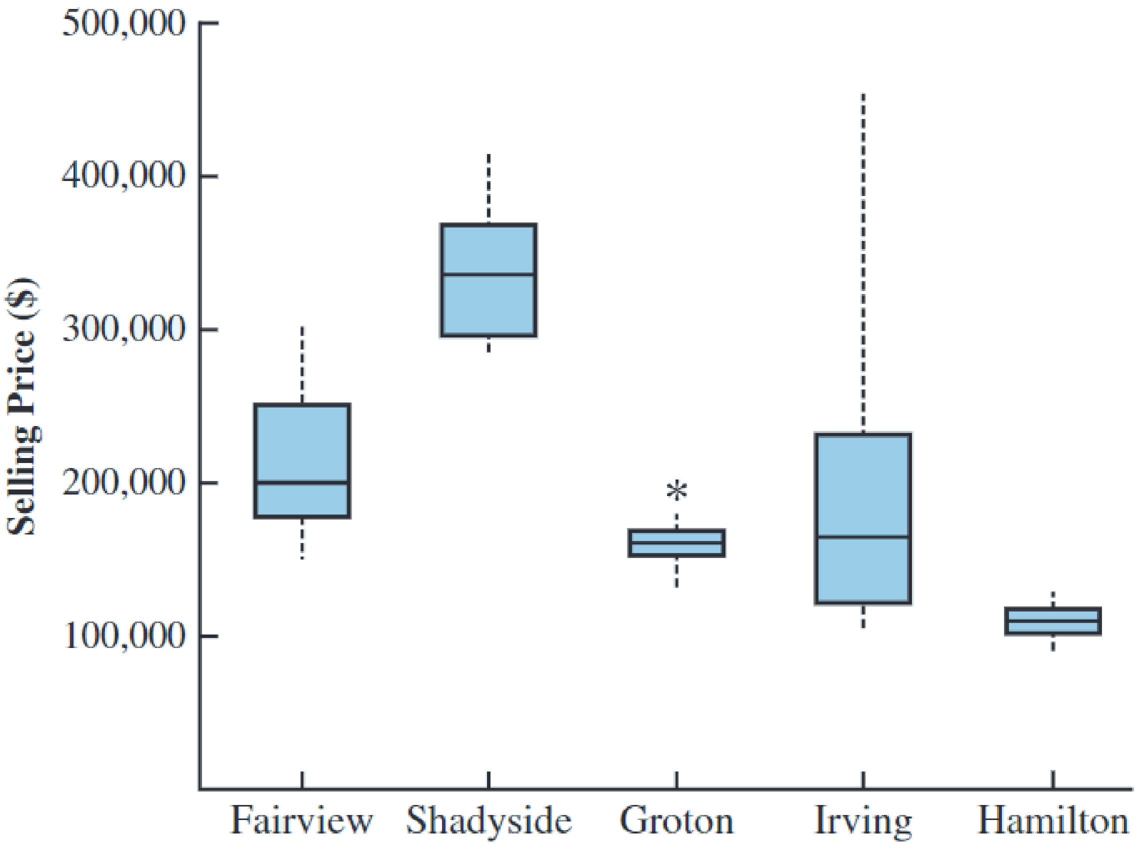
Boxplots are also very useful for comparing different data sets.

DATAfile: *homesalescomparison*

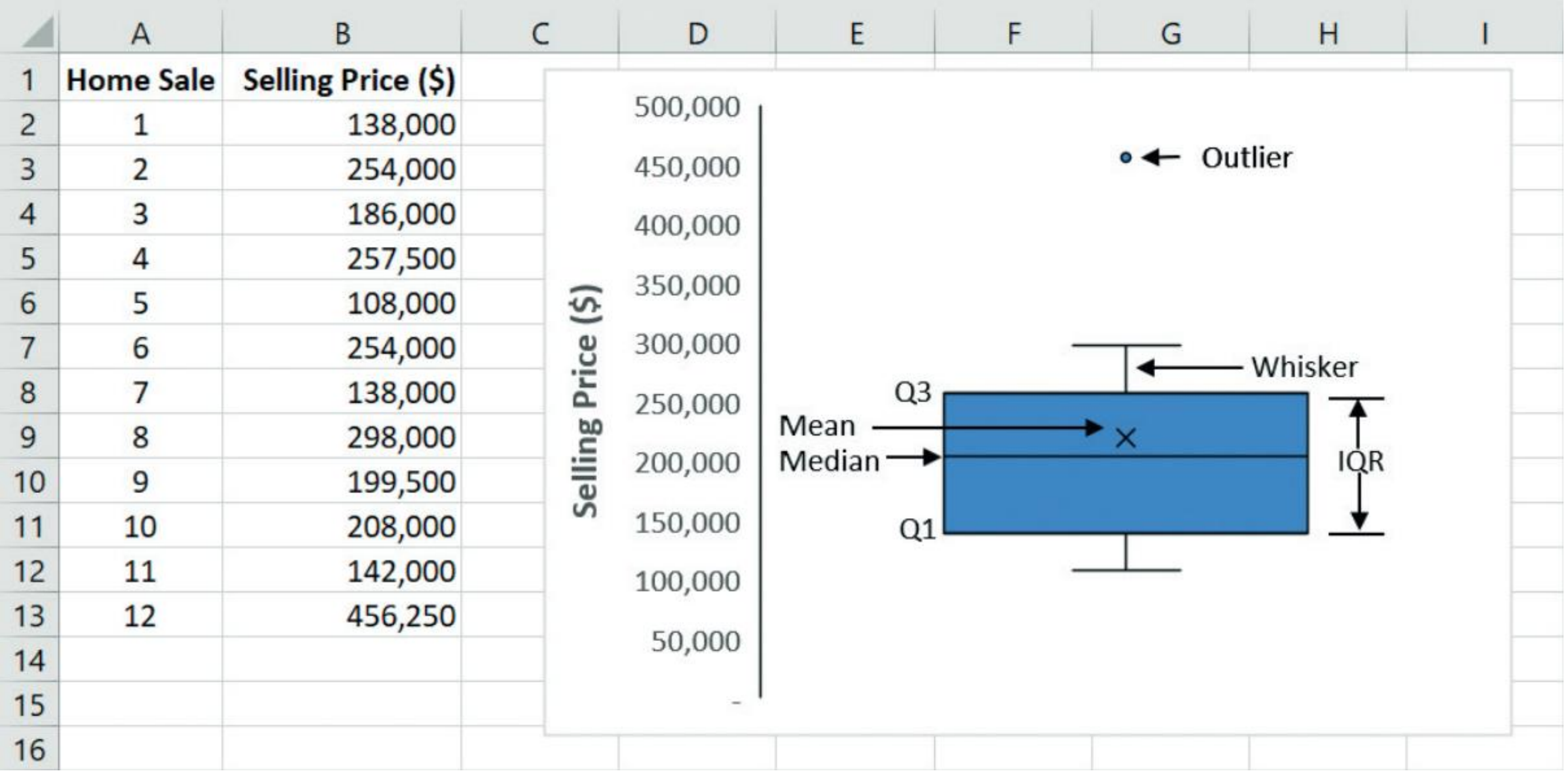
For instance, if we want to compare home sales from several communities, we could create boxplots for recent home sales in each community.

An example of such boxplots is shown to the right.

Insight is summarized in the notes.



2.7 Boxplot Created in Excel for the Home Sales Data



Home Sale	Selling Price ($)
1	138,000
2	254,000
3	186,000
4	257,500
5	108,000
6	254,000
7	138,000
8	298,000
9	199,500
10	208,000
11	142,000
12	456,250

See notes for the steps in Excel.

2.8 Scatter Charts

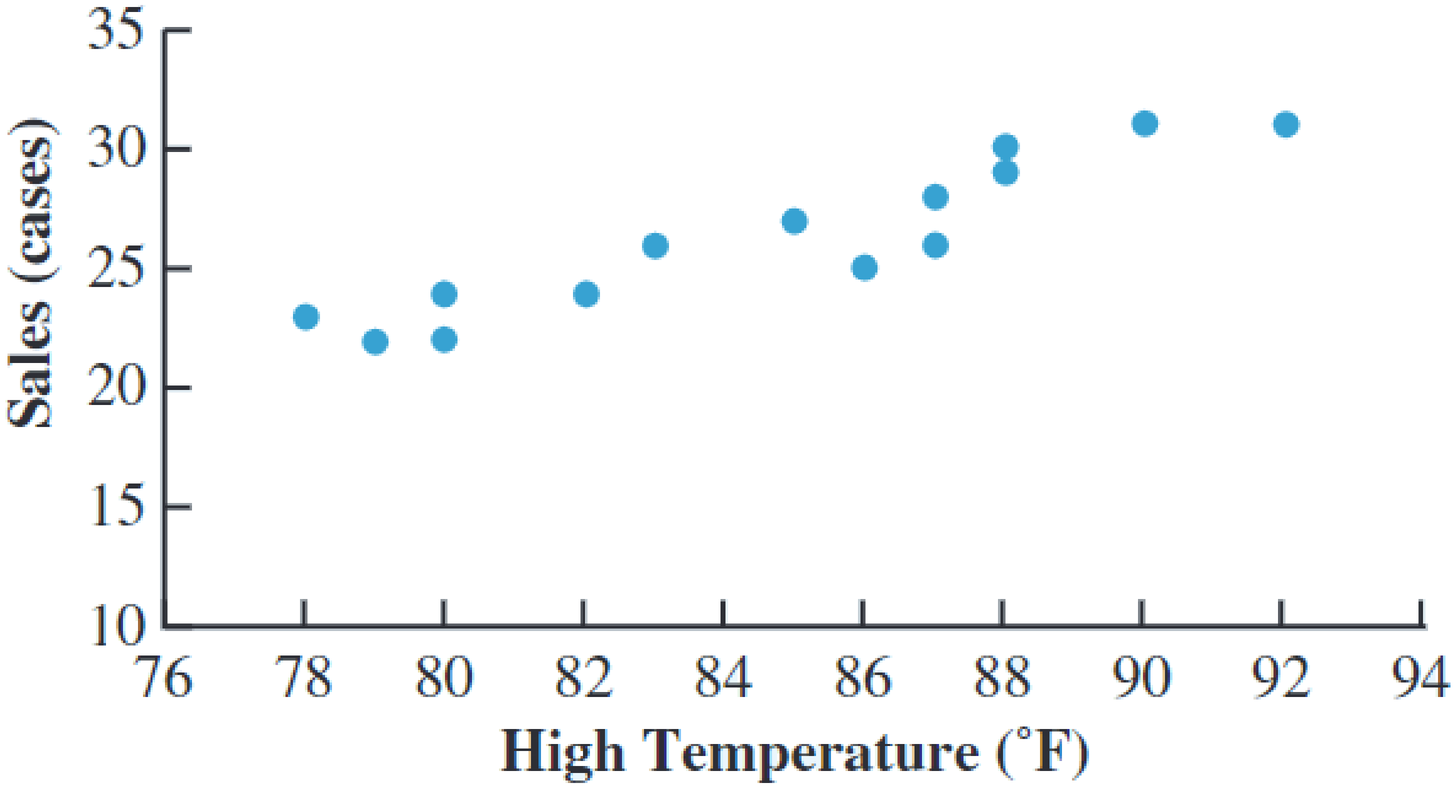
A **scatter chart** is a graph for analyzing the relationship between two variables.

DATAfile: *bottledwater*

The scatter chart for the bottled water sales at Queensland Amusement Park over 14 summer days suggests a

- positive and linear relationship

between high temperature and bottled water sales.



2.8 Covariance

Covariance is a descriptive measure of the linear association between two variables.

For a sample of size n with observations (x_i, y_i), the sample covariance is:

$$\boldsymbol{s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n - 1}}$$

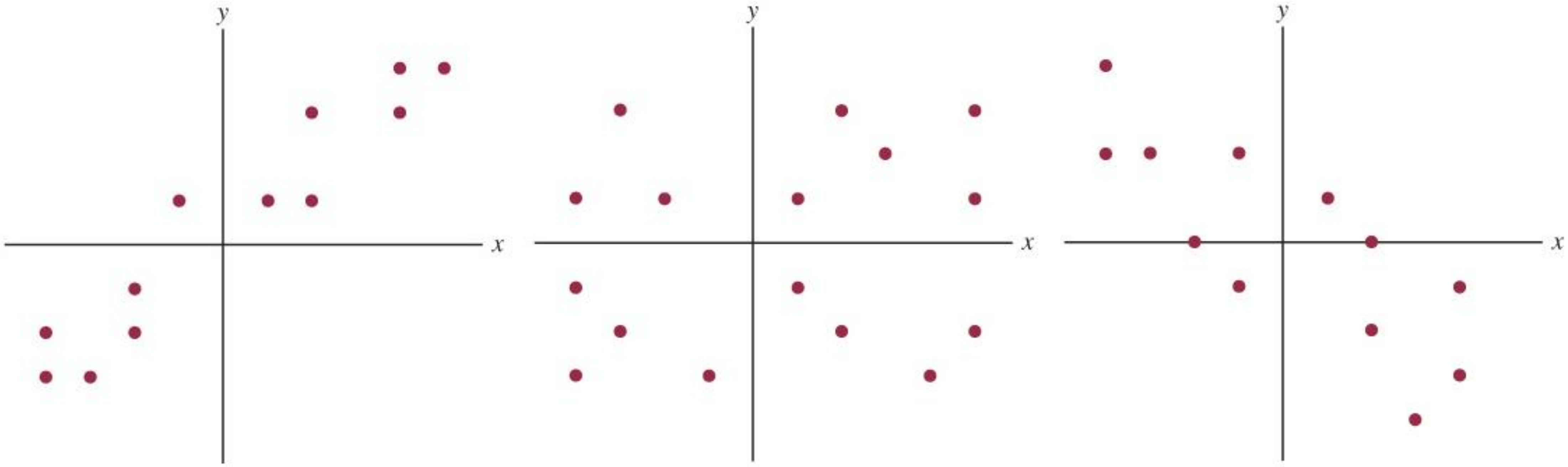
For the bottled water data, we can calculate the sample covariance between high temperatures and sales using Excel as:

s_{xy} = *COVARIANCE.S(A2:A15,B2:B15)* = 12.80

The positive covariance indicates a positive relationship between the variables.

See notes for the definition of the covariance of population data.

2.8 Relationship Between Pattern and Covariance



s_{xy} Positive:
(x and y are positively linearly related)

s_{xy} Approximately 0:
(x and y are not linearly related)

s_{xy} Negative:
(x and y are negatively linearly related)

2.8 Correlation Coefficient

One limitation of the covariance to describe the relationship between two variables is that its magnitude depends on the variables' units of measurement.

To remedy such limitation, we use the correlation coefficient to measure the strength of a relationship between two variables.

For a sample of size n, the correlation coefficient is defined as follows:

$$\boldsymbol{r_{xy} = \frac{s_{xy}}{s_x s_y}}$$

Using Excel, the sample correlation coefficient for the bottled water data is:

r_{xy} = *CORREL(A2:A15,B2:B15)* = 0.93

See notes for the definition of the correlation coefficient of population data.

2.8 Interpretation of the Correlation Coefficient

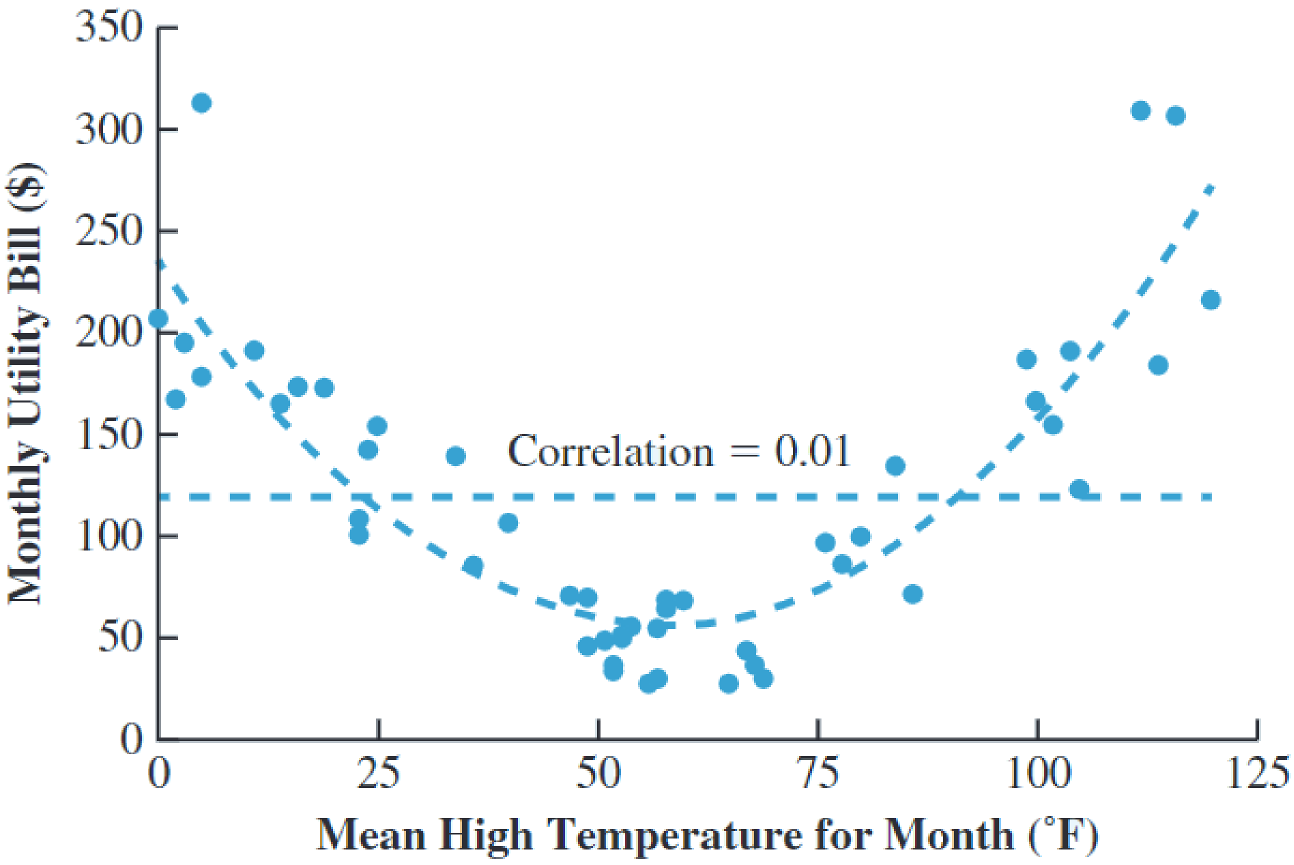
- The correlation coefficient is always between −1 and +1.
- The sign of the correlation coefficient matches the direction of the association.
- The closer to +1, the stronger the positive linear association between *x* and *y*.
- The closer to −1, the stronger the negative linear association between *x* and *y*.
- The closer to 0, the weaker the linear association between *x* and *y*.

For the bottled water data, $r = 0.93$ indicates a strong positive relationship between sales and high temperatures.

2.8 A Note on Linear Association

The correlation coefficient measures only the strength of the *linear association* between two quantitative variables.

The sample correlation coefficient for the data shown here is $r_{xy} = 0.01$ and indicates no linear relationship between the two variables, despite the strong visual evidence of a *nonlinear* relationship.



Summary

- In this chapter, we have introduced descriptive statistics to summarize data.
- We began by defining data types and data sources.
- We presented several useful functions for modifying data in Excel.
- We introduced the concept of a distribution and explained how to describe it using different interpretations of the frequency of counts and visualize it.
- We then introduced measures of location, such as mean and median, and variability, such as variance and standard deviation.
- We also presented additional measures for analyzing data distributions.
- Finally, we discussed how to visualize the relationship between two variables and how to measure their linear association using covariance and correlation coefficient.