2–2

Histograms, Frequency Polygons, and Ogives

Objective 2

Represent data in frequency distributions graphically using histograms, frequency polygons, and ogives.

After you have organized the data into a frequency distribution, you can present them in graphical form. The purpose of graphs in statistics is to convey the data to the viewers in pictorial form. It is easier for most people to comprehend the meaning of data presented graphically than data presented numerically in tables or frequency distributions. This is especially true if the users have little or no statistical knowledge.

Statistical graphs can be used to describe the data set or to analyze it. Graphs are also useful in getting the audience's attention in a publication or a speaking presentation. They can be used to discuss an issue, reinforce a critical point, or summarize a data set. They can also be used to discover a trend or pattern in a situation over a period of time.

The three most commonly used graphs in research are

1. The histogram.
2. The frequency polygon.
3. The cumulative frequency graph, or ogive (pronounced o-jive).

An example of each type of graph is shown in Figure 2–1. The data for each graph are the distribution of the miles that 20 randomly selected runners ran during a given week.

Historical Note

Karl Pearson introduced the histogram in 1891. He used it to show time concepts of various reigns of Prime Ministers.

The Histogram

The **histogram** is a graph that displays the data by using contiguous vertical bars (unless the frequency of a class is 0) of various heights to represent the frequencies of the classes.

Example 2–4

Record High Temperatures

Construct a histogram to represent the data shown for the record high temperatures for each of the 50 states (see Example 2–2).

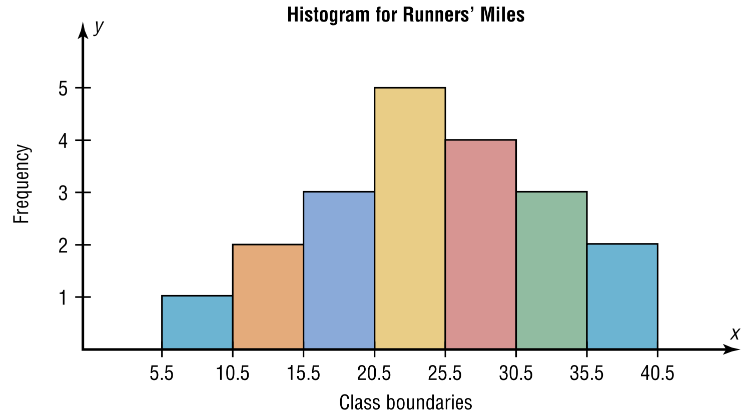
Class boundaries	Frequency
99.5–104.5	2
104.5–109.5	8
109.5–114.5	18
114.5–119.5	13
119.5–124.5	7
124.5–129.5	1
129.5–134.5	1

Solution

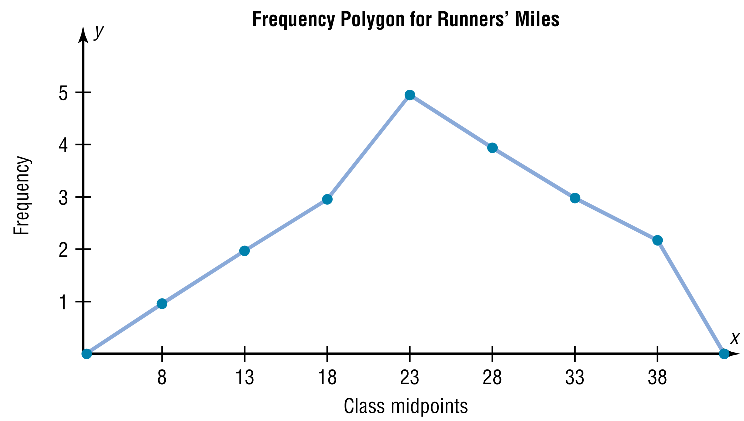
Step 1 Draw and label the x and y axes. The x axis is always the horizontal axis, and the y axis is always the vertical axis.

Figure 2–1

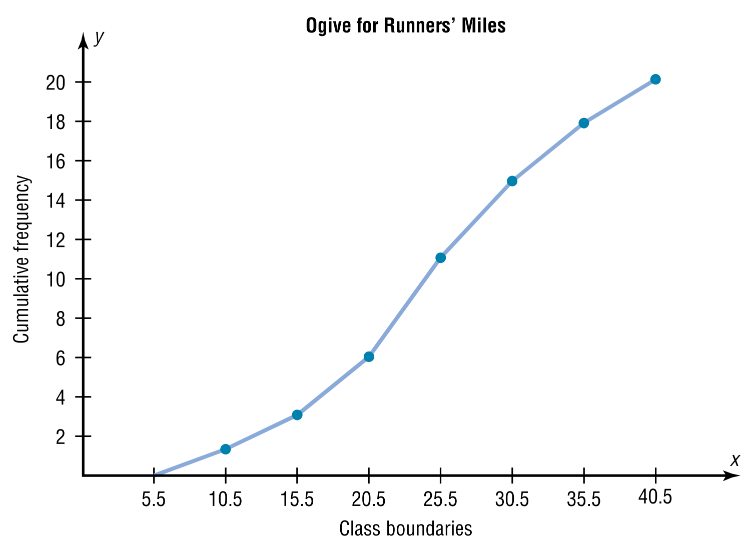
Examples of Commonly Used Graphs



(a) Histogram



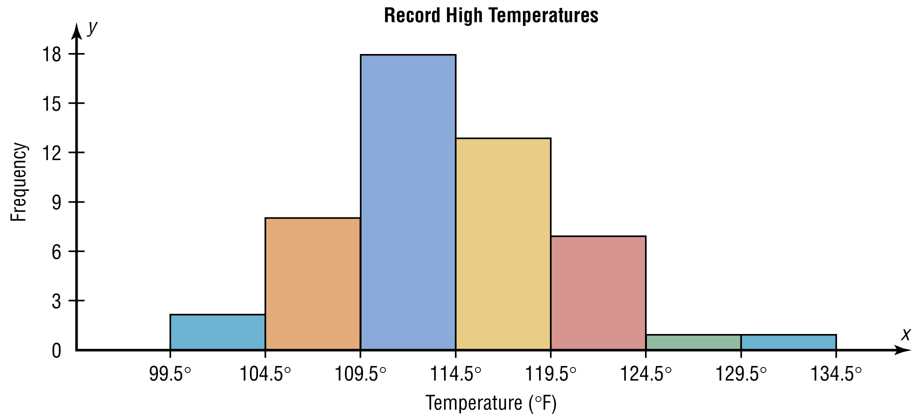
(b) Frequency polygon



(c) Cumulative frequency graph

Figure 2–2

Histogram for Example 2–4



Historical Note

Graphs originated when ancient astronomers drew the position of the stars in the heavens. Roman surveyors also used coordinates to locate landmarks on their maps.

The development of statistical graphs can be traced to William Playfair (1748–1819), an engineer and drafter who used graphs to present economic data pictorially.

Step 2 Represent the frequency on the *y* axis and the class boundaries on the *x* axis.

Step 3 Using the frequencies as the heights, draw vertical bars for each class. See Figure 2–2.

As the histogram shows, the class with the greatest number of data values (18) is 109.5–114.5, followed by 13 for 114.5–119.5. The graph also has one peak with the data clustering around it.

The Frequency Polygon

Another way to represent the same data set is by using a frequency polygon.

The **frequency polygon** is a graph that displays the data by using lines that connect points plotted for the frequencies at the midpoints of the classes. The frequencies are represented by the heights of the points.

Example 2–5 shows the procedure for constructing a frequency polygon.

Example 2–5

Record High Temperatures

Using the frequency distribution given in Example 2–4, construct a frequency polygon.

Solution

Step 1 Find the midpoints of each class. Recall that midpoints are found by adding the upper and lower boundaries and dividing by 2:

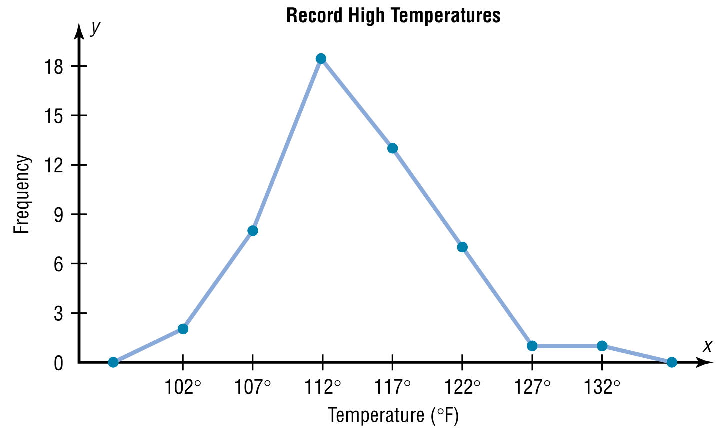
$$\frac{99.5 + 104.5}{2} = 102 \qquad \frac{104.5 + 109.5}{2} = 107$$

and so on. The midpoints are

Class boundaries	Midpoints	Frequency
99.5–104.5	102	2
104.5–109.5	107	8
109.5–114.5	112	18
114.5–119.5	117	13
119.5–124.5	122	7
124.5–129.5	127	1
129.5–134.5	132	1

Figure 2–3

Frequency Polygon for Example 2–5



Step 2 Draw the *x* and *y* axes. Label the *x* axis with the midpoint of each class, and then use a suitable scale on the *y* axis for the frequencies.

Step 3 Using the midpoints for the *x* values and the frequencies as the *y* values, plot the points.

Step 4 Connect adjacent points with line segments. Draw a line back to the *x* axis at the beginning and end of the graph, at the same distance that the previous and next midpoints would be located, as shown in Figure 2–3.

The frequency polygon and the histogram are two different ways to represent the same data set. The choice of which one to use is left to the discretion of the researcher.

The Ogive

The third type of graph that can be used represents the cumulative frequencies for the classes. This type of graph is called the *cumulative frequency graph,* or *ogive.* The **cumulative frequency** is the sum of the frequencies accumulated up to the upper boundary of a class in the distribution.

The **ogive** is a graph that represents the cumulative frequencies for the classes in a frequency distribution.

Example 2–6 shows the procedure for constructing an ogive.

Example 2–6

Record High Temperatures

Construct an ogive for the frequency distribution described in Example 2–4.

Solution

Step 1 Find the cumulative frequency for each class.

	Cumulative frequency
Less than 99.5	0
Less than 104.5	2
Less than 109.5	10
Less than 114.5	28
Less than 119.5	41
Less than 124.5	48
Less than 129.5	49
Less than 134.5	50

Figure 2–4

Plotting the Cumulative Frequency for Example 2–6

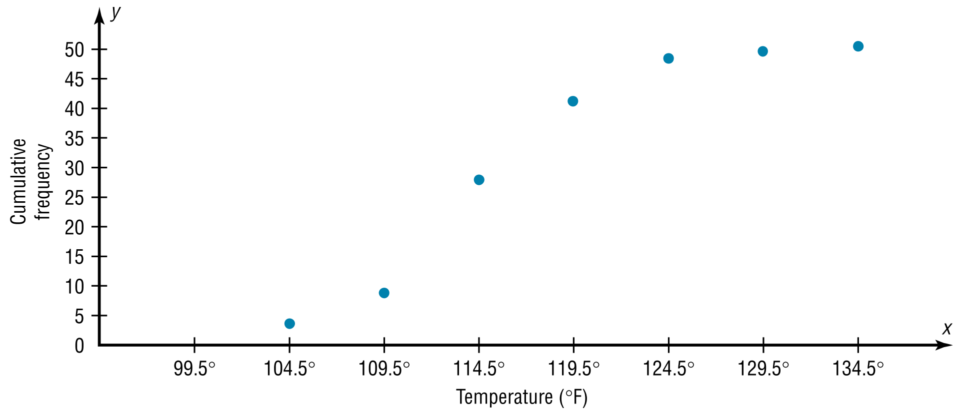
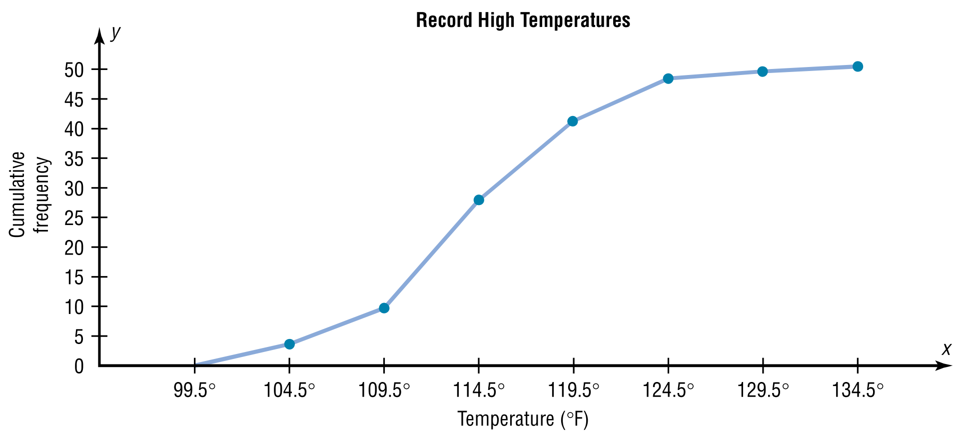


Figure 2–5

Ogive for Example 2–6



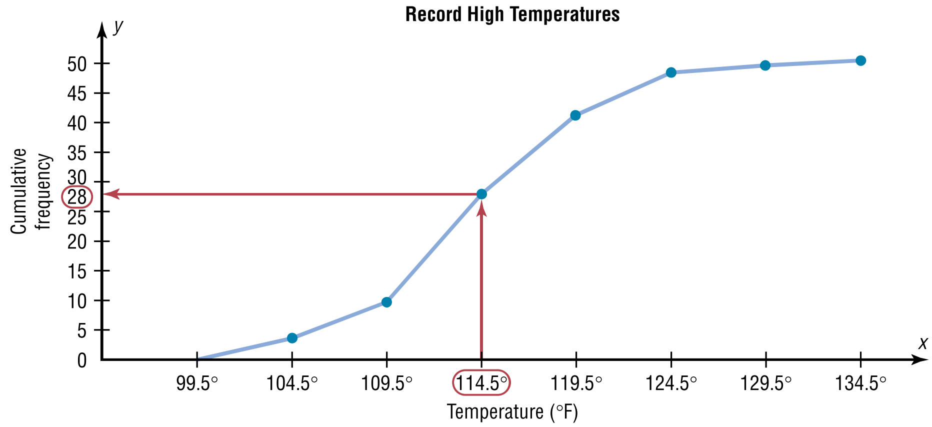
Step 2 Draw the *x* and *y* axes. Label the *x* axis with the class boundaries. Use an appropriate scale for the *y* axis to represent the cumulative frequencies. (Depending on the numbers in the cumulative frequency columns, scales such as 0, 1, 2, 3, . . . , or 5, 10, 15, 20, . . . , or 1000, 2000, 3000, . . . can be used. Do *not* label the *y* axis with the numbers in the cumulative frequency column.) In this example, a scale of 0, 5, 10, 15, . . . will be used.

Step 3 Plot the cumulative frequency at each upper class boundary, as shown in Figure 2–4. Upper boundaries are used since the cumulative frequencies represent the number of data values accumulated up to the upper boundary of each class.

Step 4 Starting with the first upper class boundary, 104.5, connect adjacent points with line segments, as shown in Figure 2–5. Then extend the graph to the first lower class boundary, 99.5, on the *x* axis.

Cumulative frequency graphs are used to visually represent how many values are below a certain upper class boundary. For example, to find out how many record high temperatures are less than 114.5°F, locate 114.5°F on the *x* axis, draw a vertical line up until it intersects the graph, and then draw a horizontal line at that point to the *y* axis. The *y* axis value is 28, as shown in Figure 2–6.

Figure 2–6
Finding a Specific Cumulative Frequency



The steps for drawing these three types of graphs are shown in the following Procedure Table.

Unusual Stat

Twenty-two percent of Americans sleep 6 hours a day or fewer.

Procedure Table

Constructing Statistical Graphs

Step 1 Draw and label the x and y axes.

Step 2 Choose a suitable scale for the frequencies or cumulative frequencies, and label it on the y axis.

Step 3 Represent the class boundaries for the histogram or ogive, or the midpoint for the frequency polygon, on the x axis.

Step 4 Plot the points and then draw the bars or lines.

Relative Frequency Graphs

The histogram, the frequency polygon, and the ogive shown previously were constructed by using frequencies in terms of the raw data. These distributions can be converted to distributions using *proportions* instead of raw data as frequencies. These types of graphs are called **relative frequency graphs.**

Graphs of relative frequencies instead of frequencies are used when the proportion of data values that fall into a given class is more important than the actual number of data values that fall into that class. For example, if you wanted to compare the age distribution of adults in Philadelphia, Pennsylvania, with the age distribution of adults of Erie, Pennsylvania, you would use relative frequency distributions. The reason is that since the population of Philadelphia is 1,478,002 and the population of Erie is 105,270, the bars using the actual data values for Philadelphia would be much taller than those for the same classes for Erie.

To convert a frequency into a proportion or relative frequency, divide the frequency for each class by the total of the frequencies. The sum of the relative frequencies will always be 1. These graphs are similar to the ones that use raw data as frequencies, but the values on the y axis are in terms of proportions. Example 2–7 shows the three types of relative frequency graphs.

Example 2–7

Miles Run per Week

Construct a histogram, frequency polygon, and ogive using relative frequencies for the distribution (shown here) of the miles that 20 randomly selected runners ran during a given week.

Class boundaries	Frequency
5.5–10.5	1
10.5–15.5	2
15.5–20.5	3
20.5–25.5	5
25.5–30.5	4
30.5–35.5	3
35.5–40.5	2
	20

Solution

Step 1 Convert each frequency to a proportion or relative frequency by dividing the frequency for each class by the total number of observations.

For class 5.5–10.5, the relative frequency is $\frac{1}{20} = 0.05$; for class 10.5–15.5, the relative frequency is $\frac{2}{20} = 0.10$; for class 15.5–20.5, the relative frequency is $\frac{3}{20} = 0.15$; and so on.

Place these values in the column labeled Relative frequency.

Class boundaries	Midpoints	Relative frequency
5.5–10.5	8	0.05
10.5–15.5	13	0.10
15.5–20.5	18	0.15
20.5–25.5	23	0.25
25.5–30.5	28	0.20
30.5–35.5	33	0.15
35.5–40.5	38	0.10
		1.00

Step 2 Find the cumulative relative frequencies. To do this, add the frequency in each class to the total frequency of the preceding class. In this case, $0 + 0.05 = 0.05$, $0.05 + 0.10 = 0.15$, $0.15 + 0.15 = 0.30$, $0.30 + 0.25 = 0.55$, etc. Place these values in the column labeled Cumulative relative frequency.

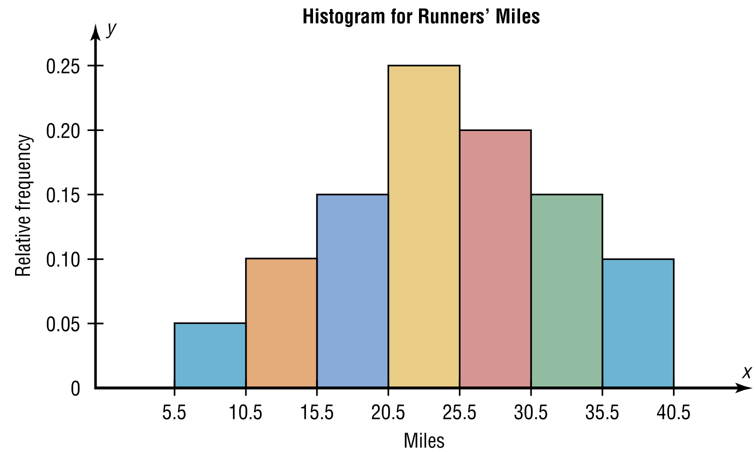
An alternative method would be to find the cumulative frequencies and then convert each one to a relative frequency.

	Cumulative frequency	Cumulative relative frequency
Less than 5.5	0	0.00
Less than 10.5	1	0.05
Less than 15.5	3	0.15
Less than 20.5	6	0.30
Less than 25.5	11	0.55
Less than 30.5	15	0.75
Less than 35.5	18	0.90
Less than 40.5	20	1.00

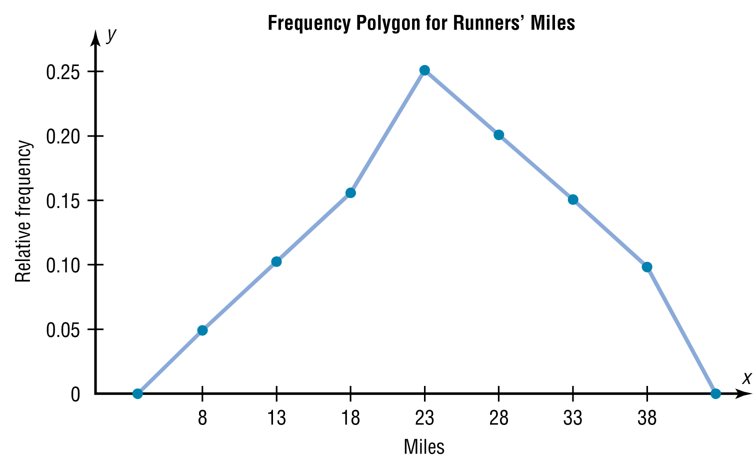
Step 3 Draw each graph as shown in Figure 2–7. For the histogram and ogive, use the class boundaries along the *x* axis. For the frequency polygon, use the midpoints on the *x* axis. The scale on the *y* axis uses proportions.

Figure 2–7

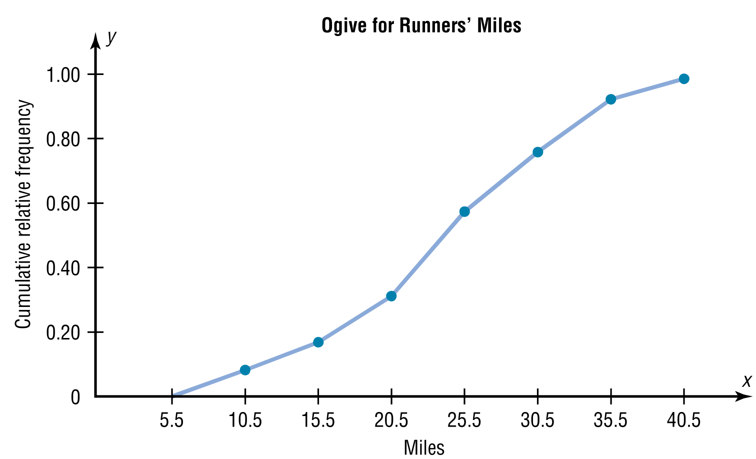
Graphs for Example 2–7



(a) Histogram



(b) Frequency polygon



(c) Ogive

Distribution Shapes

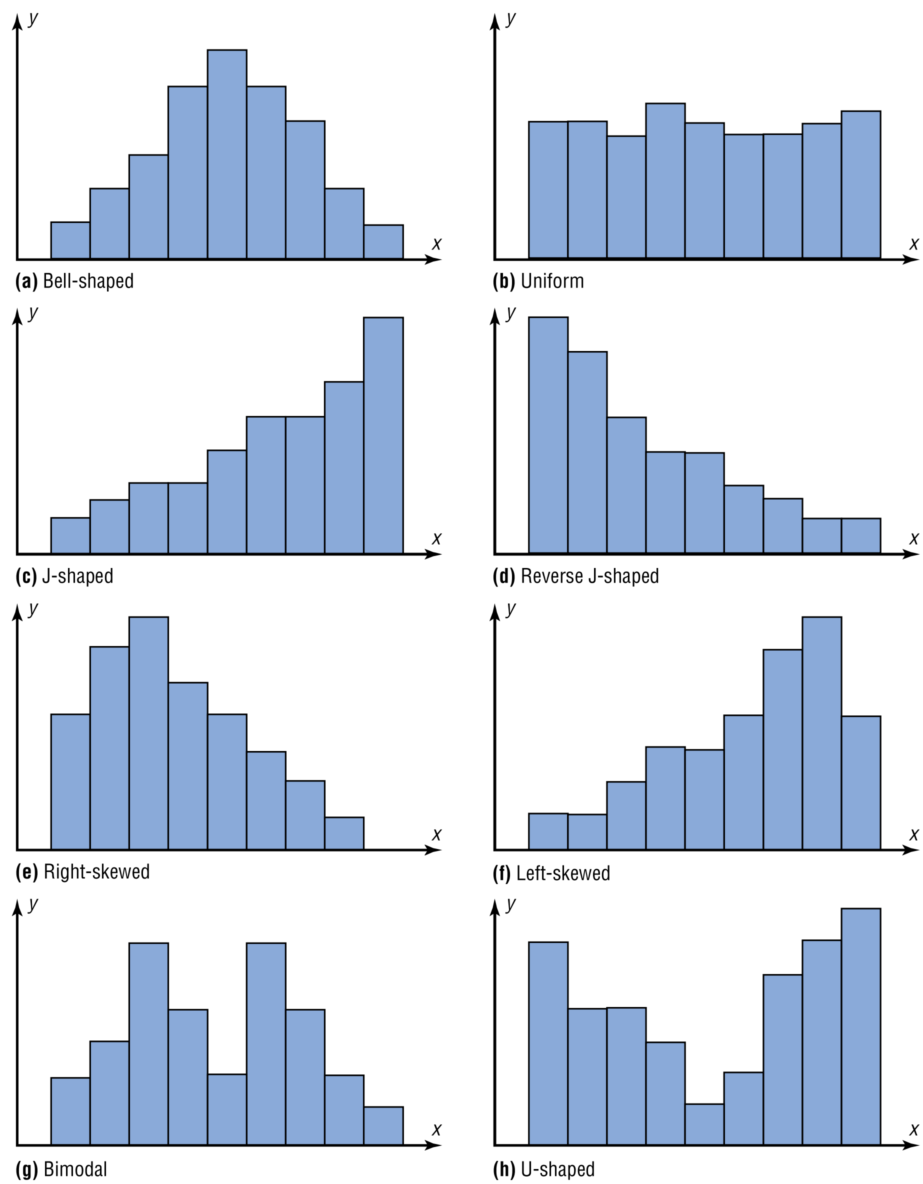
When one is describing data, it is important to be able to recognize the shapes of the distribution values. In later chapters you will see that the shape of a distribution also determines the appropriate statistical methods used to analyze the data.

A distribution can have many shapes, and one method of analyzing a distribution is to draw a histogram or frequency polygon for the distribution. Several of the most common shapes are shown in Figure 2–8: *the bell-shaped or mound-shaped, the uniform-shaped, the J-shaped, the reverse J-shaped, the positively or right-skewed shape, the negatively or left-skewed shape, the bimodal-shaped, and the U-shaped.*

Distributions are most often not perfectly shaped, so it is not necessary to have an exact shape but rather to identify an overall pattern.

A *bell-shaped distribution* shown in Figure 2–8(a) has a single peak and tapers off at either end. It is approximately symmetric; i.e., it is roughly the same on both sides of a line running through the center.

Figure 2–8
Distribution Shapes



A *uniform distribution* is basically flat or rectangular. See Figure 2–8(b).

A *J-shaped distribution* is shown in Figure 2–8(c), and it has a few data values on the left side and increases as one moves to the right. A *reverse J-shaped distribution* is the opposite of the J-shaped distribution. See Figure 2–8(d).

When the peak of a distribution is to the left and the data values taper off to the right, a distribution is said to be *positively or right-skewed.* See Figure 2–8(e). When the data values are clustered to the right and taper off to the left, a distribution is said to be *negatively or left-skewed.* See Figure 2–8(f). Skewness will be explained in detail in Chapter 3. Distributions with one peak, such as those shown in Figure 2–8(a), (e), and (f), are said to be *unimodal.* (The highest peak of a distribution indicates where the mode of the data values is. The mode is the data value that occurs more often than any other data value. Modes are explained in Chapter 3.) When a distribution has two peaks of the same height, it is said to be *bimodal.* See Figure 2–8(g). Finally, the graph shown in Figure 2–8(h) is a *U-shaped* distribution.

Distributions can have other shapes in addition to the ones shown here; however, these are some of the more common ones that you will encounter in analyzing data.

When you are analyzing histograms and frequency polygons, look at the shape of the curve. For example, does it have one peak or two peaks? Is it relatively flat, or is it U-shaped? Are the data values spread out on the graph, or are they clustered around the center? Are there data values in the extreme ends? These may be *outliers.* (See Section 3–3 for an explanation of outliers.) Are there any gaps in the histogram, or does the frequency polygon touch the *x* axis somewhere other than at the ends? Finally, are the data clustered at one end or the other, indicating a *skewed distribution*?

For example, the histogram for the record high temperatures shown in Figure 2–2 shows a single peaked distribution, with the class 109.5–114.5 containing the largest number of temperatures. The distribution has no gaps, and there are fewer temperatures in the highest class than in the lowest class.

Applying the Concepts 2–2

Selling Real Estate

Assume you are a realtor in Bradenton, Florida. You have recently obtained a listing of the selling prices of the homes that have sold in that area in the last 6 months. You wish to organize those data so you will be able to provide potential buyers with useful information. Use the following data to create a histogram, frequency polygon, and cumulative frequency polygon.

142,000	127,000	99,600	162,000	89,000	93,000	99,500
73,800	135,000	119,500	67,900	156,300	104,500	108,650
123,000	91,000	205,000	110,000	156,300	104,000	133,900
179,000	112,000	147,000	321,550	87,900	88,400	180,000
159,400	205,300	144,400	163,000	96,000	81,000	131,000
114,000	119,600	93,000	123,000	187,000	96,000	80,000
231,000	189,500	177,600	83,400	77,000	132,300	166,000

1. What questions could be answered more easily by looking at the histogram rather than the listing of home prices?
2. What different questions could be answered more easily by looking at the frequency polygon rather than the listing of home prices?
3. What different questions could be answered more easily by looking at the cumulative frequency polygon rather than the listing of home prices?
4. Are there any extremely large or extremely small data values compared to the other data values?
5. Which graph displays these extremes the best?
6. Is the distribution skewed?

See page 101 for the answers.

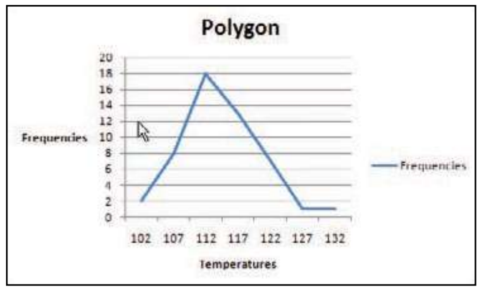
3. Select Edit from the Horizontal Axis Labels and highlight the midpoints from column A, then click [OK].
4. Click [OK] on the Select Data Source box.

Inserting Labels on the Axes

1. Click the mouse on any region of the graph.
2. Select Chart Tools and then Layout on the toolbar.
3. Select Axis Titles to open the horizontal and vertical axis text boxes. Then manually type in labels for the axes.

Changing the Title

1. Select Chart Tools, Layout from the toolbar.
2. Select Chart Title.
3. Choose one of the options from the Chart Title menu and edit.

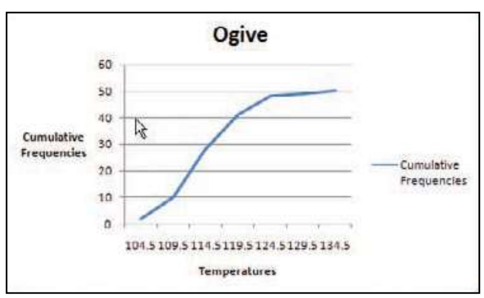


Constructing an Ogive

To create an ogive, you can use the upper class boundaries (horizontal axis) and cumulative frequencies (vertical axis) from the frequency distribution.

1. Type the upper class boundaries and cumulative frequencies into adjacent columns of an Excel worksheet.
2. Highlight the cumulative frequencies (including the label) and select the Insert tab from the toolbar.
3. Select Line Chart, then the 2-D Line option.

As with the frequency polygon, you can insert labels on the axes and a chart title for the ogive.

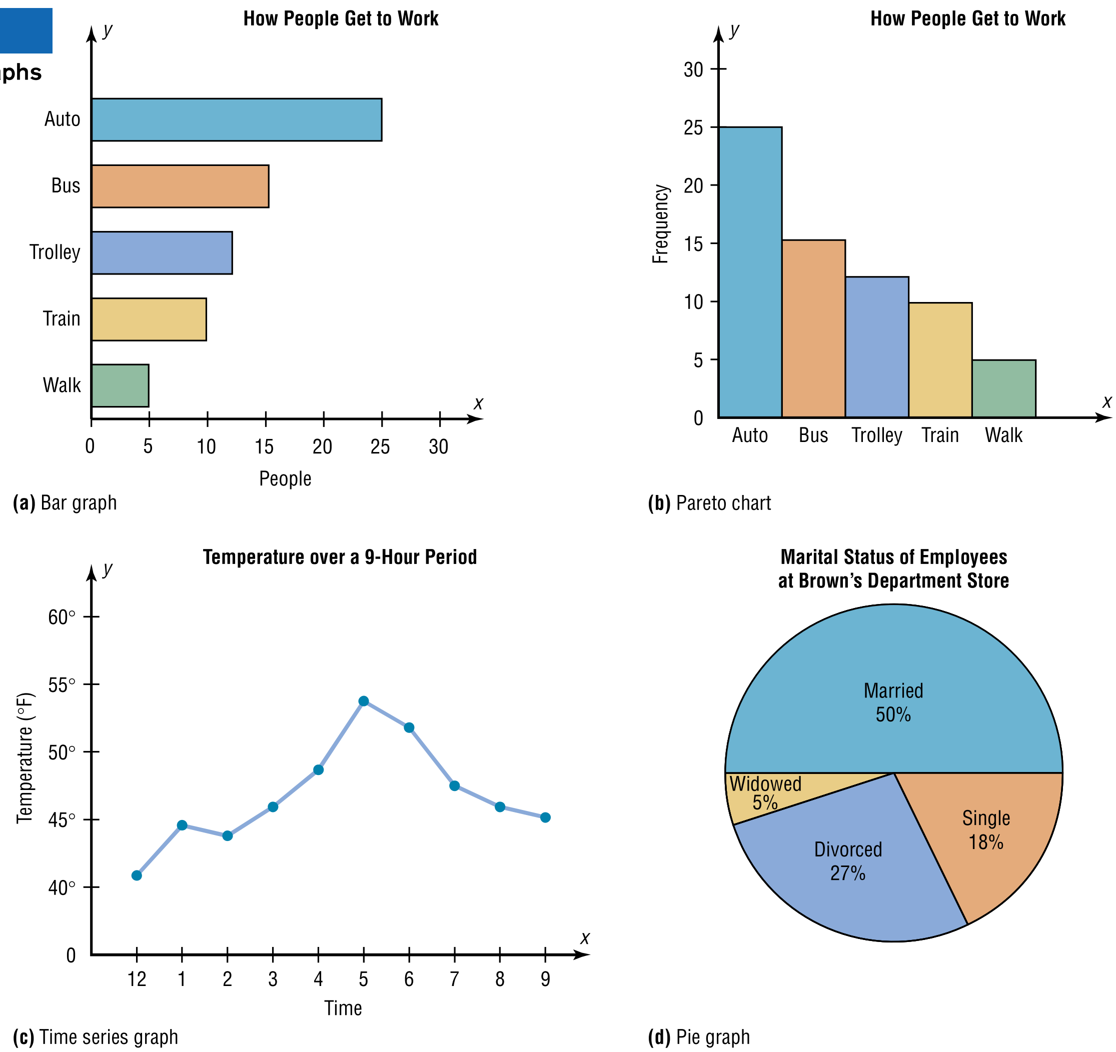


2–3 Other Types of Graphs

In addition to the histogram, the frequency polygon, and the ogive, several other types of graphs are often used in statistics. They are the bar graph, Pareto chart, time series graph, and pie graph. Figure 2–9 shows an example of each type of graph.

Figure 2–9

Other Types of Graphs Used in Statistics



Objective 3

Represent data using bar graphs, Pareto charts, time series graphs, and pie graphs.

Bar Graphs

When the data are qualitative or categorical, bar graphs can be used to represent the data. A bar graph can be drawn using either horizontal or vertical bars.

A **bar graph** represents the data by using vertical or horizontal bars whose heights or lengths represent the frequencies of the data.

Example 2–8 College Spending for First-Year Students

The table shows the average money spent by first-year college students. Draw a horizontal and vertical bar graph for the data.

Electronics	$728
Dorm decor	344
Clothing	141
Shoes	72

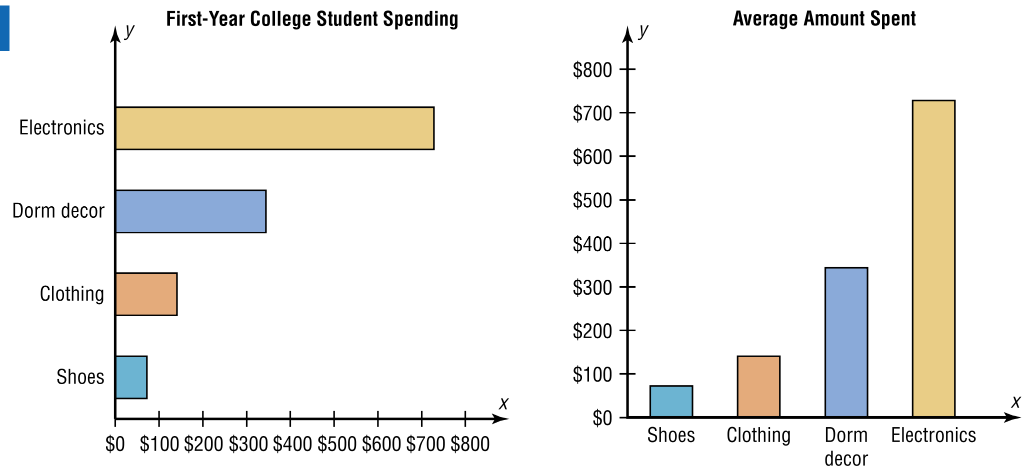
Source: The National Retail Federation.

Solution

1. Draw and label the *x* and *y* axes. For the horizontal bar graph place the frequency scale on the *x* axis, and for the vertical bar graph place the frequency scale on the *y* axis.
2. Draw the bars corresponding to the frequencies. See Figure 2–10.

Figure 2–10

Bar Graphs for Example 2–8



The graphs show that first-year college students spend the most on electronic equipment including computers.

Pareto Charts

When the variable displayed on the horizontal axis is qualitative or categorical, a *Pareto chart* can also be used to represent the data.

A **Pareto chart** is used to represent a frequency distribution for a categorical variable, and the frequencies are displayed by the heights of vertical bars, which are arranged in order from highest to lowest.

Example 2–9

Homeless People

The data shown here consist of the number of homeless people for a sample of selected cities. Construct and analyze a Pareto chart for the data.

City	Number
Atlanta	6832
Baltimore	2904
Chicago	6680
St. Louis	1485
Washington	5518

Source: U.S. Department of Housing and Urban Development.

Historical Note

Vilfredo Pareto (1848–1923) was an Italian scholar who developed theories in economics, statistics, and the social sciences. His contributions to statistics include the development of a mathematical function used in economics. This function has many statistical applications and is called the Pareto distribution. In addition, he researched income distribution, and his findings became known as Pareto's law.

Solution

Step 1 Arrange the data from the largest to smallest according to frequency.

City	Number
Atlanta	6832
Chicago	6680
Washington	5518
Baltimore	2904
St. Louis	1485

Step 2 Draw and label the x and y axes.

Step 3 Draw the bars corresponding to the frequencies. See Figure 2–11.

The graph shows that the number of homeless people is about the same for Atlanta and Chicago and a lot less for Baltimore and St. Louis.

Suggestions for Drawing Pareto Charts

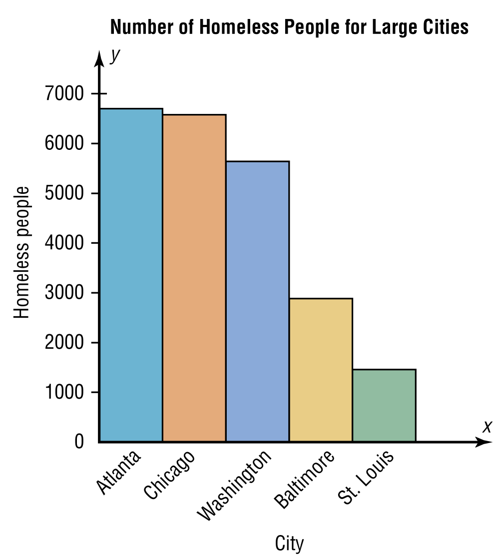
1. Make the bars the same width.
2. Arrange the data from largest to smallest according to frequency.
3. Make the units that are used for the frequency equal in size.

When you analyze a Pareto chart, make comparisons by looking at the heights of the bars.

The Time Series Graph

When data are collected over a period of time, they can be represented by a time series graph.

Figure 2–11
Pareto Chart for Example 2–9



A **time series graph** represents data that occur over a specific period of time.

Example 2–10 shows the procedure for constructing a time series graph.

Example 2–10

Workplace Homicides

The number of homicides that occurred in the workplace for the years 2003 to 2008 is shown. Draw and analyze a time series graph for the data.

Year	’03	’04	’05	’06	’07	’08
Number	632	559	567	540	628	517

Source: Bureau of Labor Statistics.

Solution

Step 1 Draw and label the x and y axes.

Step 2 Label the x axis for years and the y axis for the number.

Step 3 Plot each point according to the table.

Step 4 Draw line segments connecting adjacent points. Do not try to fit a smooth curve through the data points. See Figure 2–12.

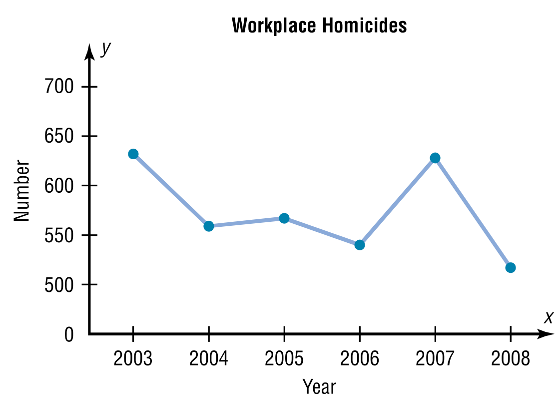
There was a slight decrease in the years ’04, ’05, and ’06, compared to ’03, and again an increase in ’07. The largest decrease occurred in ’08.

Historical Note

Time series graphs are over 1000 years old. The first ones were used to chart the movements of the planets and the sun.

Figure 2–12

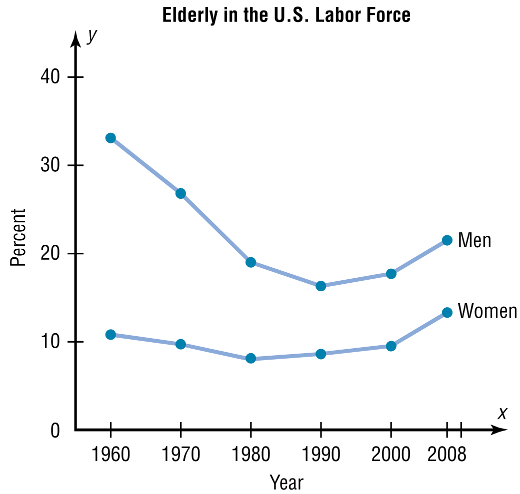
Time Series Graph for Example 2–10



When you analyze a time series graph, look for a trend or pattern that occurs over the time period. For example, is the line ascending (indicating an increase over time) or descending (indicating a decrease over time)? Another thing to look for is the slope, or steepness, of the line. A line that is steep over a specific time period indicates a rapid increase or decrease over that period.

Figure 2–13

Two Time Series Graphs for Comparison



Source: Bureau of Census, U.S. Department of Commerce.

Two or more data sets can be compared on the same graph called a *compound time series graph* if two or more lines are used, as shown in Figure 2–13. This graph shows the percentage of elderly males and females in the United States labor force from 1960 to 2008. It shows that the percent of elderly men decreased significantly from 1960 to 1990 and then increased slightly after that. For the elderly females, the percent decreased slightly from 1960 to 1980 and then increased from 1980 to 2008.

The Pie Graph

Pie graphs are used extensively in statistics. The purpose of the pie graph is to show the relationship of the parts to the whole by visually comparing the sizes of the sections. Percentages or proportions can be used. The variable is nominal or categorical.

A **pie graph** is a circle that is divided into sections or wedges according to the percentage of frequencies in each category of the distribution.

Example 2–11 shows the procedure for constructing a pie graph.

Example 2–11

Super Bowl Snack Foods

This frequency distribution shows the number of pounds of each snack food eaten during the Super Bowl. Construct a pie graph for the data.

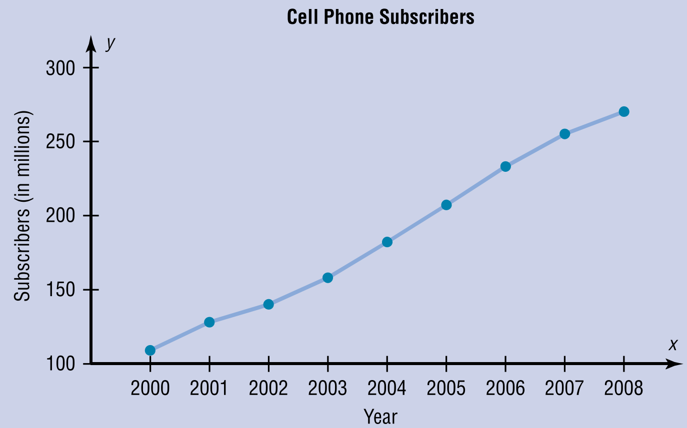
Snack	Pounds (frequency)
Potato chips	11.2 million
Tortilla chips	8.2 million
Pretzels	4.3 million
Popcorn	3.8 million
Snack nuts	2.5 million
	Total $n = 30.0$ million

Source: *USA TODAY Weekend.*

Speaking of Statistics

Cell Phone Usage

The graph shows the estimated number (in millions) of cell phone subscribers since 2000. How do you think the growth will affect our way of living? For example, emergencies can be handled faster since people are using their cell phones to call 911.



Source: The World Almanac and Book of Facts 2010.

Solution

Step 1 Since there are 360° in a circle, the frequency for each class must be converted into a proportional part of the circle. This conversion is done by using the formula

$$\text{Degrees} = \frac{f}{n} \cdot 360°$$

where f = frequency for each class and n = sum of the frequencies. Hence, the following conversions are obtained. The degrees should sum to 360°.*

Potato chips	$\frac{11.2}{30} \cdot 360° =$	134°
Tortilla chips	$\frac{8.2}{30} \cdot 360° =$	98°
Pretzels	$\frac{4.3}{30} \cdot 360° =$	52°
Popcorn	$\frac{3.8}{30} \cdot 360° =$	46°
Snack nuts	$\frac{2.5}{30} \cdot 360° =$	30°
Total		360°

Step 2 Each frequency must also be converted to a percentage. Recall from Example 2–1 that this conversion is done by using the formula

$$\% = \frac{f}{n} \cdot 100$$

Hence, the following percentages are obtained. The percentages should sum to 100%.†

Potato chips	$\frac{11.2}{30} \cdot 100 = 37.3\%$
Tortilla chips	$\frac{8.2}{30} \cdot 100 = 27.3\%$

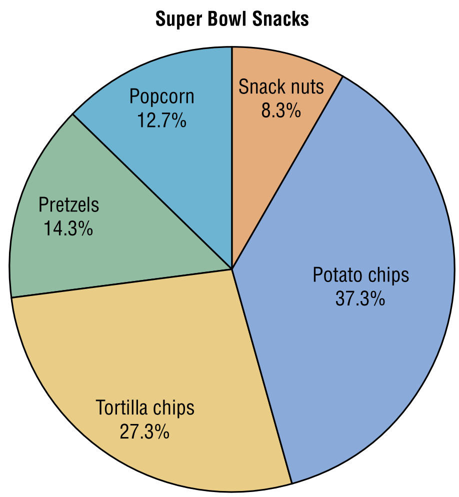
**Note:* The degrees column does not always sum to 360° due to rounding.
†*Note:* The percent column does not always sum to 100% due to rounding.

Pretzels	$\frac{4.3}{30} \cdot 100 = 14.3\%$
Popcorn	$\frac{3.8}{30} \cdot 100 = 12.7\%$
Snack nuts	$\frac{2.5}{30} \cdot 100 = 8.3\%$
Total	99.9%

Step 3 Next, using a protractor and a compass, draw the graph using the appropriate degree measures found in step 1, and label each section with the name and percentages, as shown in Figure 2–14.

Figure 2–14

Pie Graph for Example 2–11



Example 2–12

Construct a pie graph showing the blood types of the army inductees described in Example 2–1. The frequency distribution is repeated here.

Class	Frequency	Percent
A	5	20
B	7	28
O	9	36
AB	4	16
	25	100

Solution

Step 1 Find the number of degrees for each class, using the formula

$$\text{Degrees} = \frac{f}{n} \cdot 360°$$

For each class, then, the following results are obtained.

A	$\frac{5}{25} \cdot 360° = 72°$
B	$\frac{7}{25} \cdot 360° = 100.8°$

O $\frac{9}{25} \cdot 360° = 129.6°$

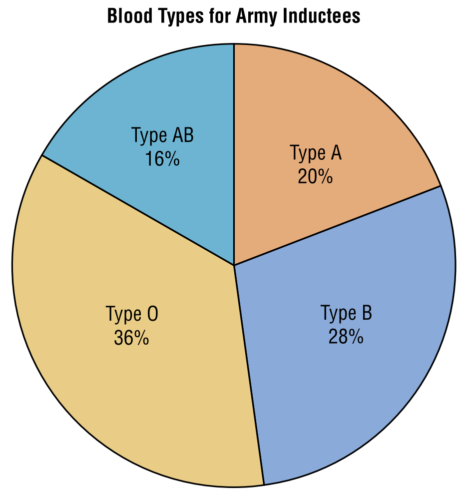
AB $\frac{4}{25} \cdot 360° = 57.6°$

Step 2 Find the percentages. (This was already done in Example 2–1.)

Step 3 Using a protractor, graph each section and write its name and corresponding percentage, as shown in Figure 2–15.

Figure 2–15

Pie Graph for Example 2–12



The graph in Figure 2–15 shows that in this case the most common blood type is type O.

To analyze the nature of the data shown in the pie graph, look at the size of the sections in the pie graph. For example, are any sections relatively large compared to the rest?

Figure 2–15 shows that among the inductees, type O blood is more prevalent than any other type. People who have type AB blood are in the minority. More than twice as many people have type O blood as type AB.

Misleading Graphs

Graphs give a visual representation that enables readers to analyze and interpret data more easily than they could simply by looking at numbers. However, inappropriately drawn graphs can misrepresent the data and lead the reader to false conclusions. For example, a car manufacturer's ad stated that 98% of the vehicles it had sold in the past 10 years were still on the road. The ad then showed a graph similar to the one in Figure 2–16. The graph shows the percentage of the manufacturer's automobiles still on the road and the percentage of its competitors' automobiles still on the road. Is there a large difference? Not necessarily.

Notice the scale on the vertical axis in Figure 2–16. It has been cut off (or truncated) and starts at 95%. When the graph is redrawn using a scale that goes from 0 to 100%, as in Figure 2–17, there is hardly a noticeable difference in the percentages. Thus, changing the units at the starting point on the y axis can convey a very different visual representation of the data.

Figure 2–16

Graph of Automaker's Claim Using a Scale from 95 to 100%

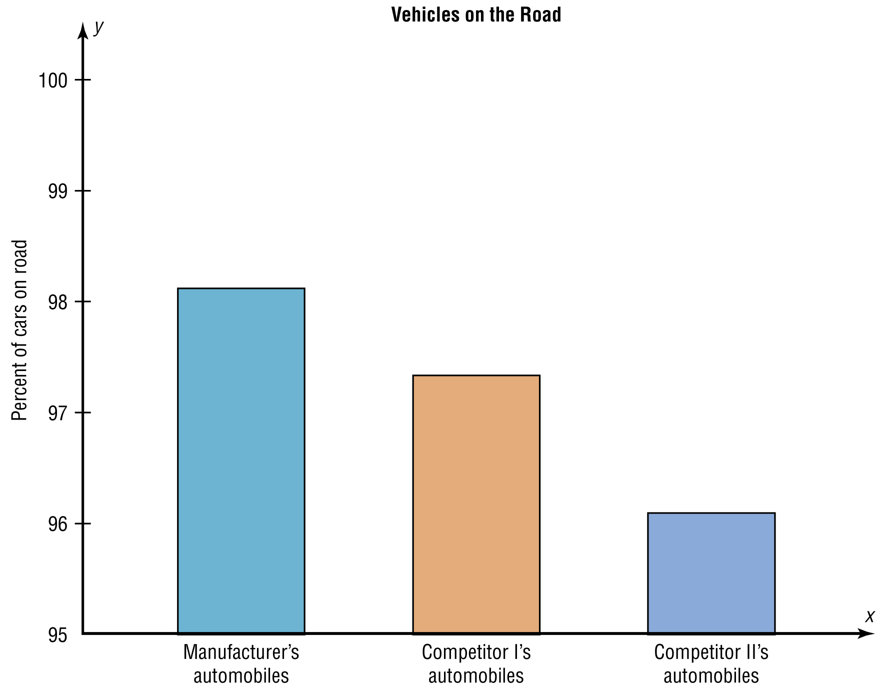
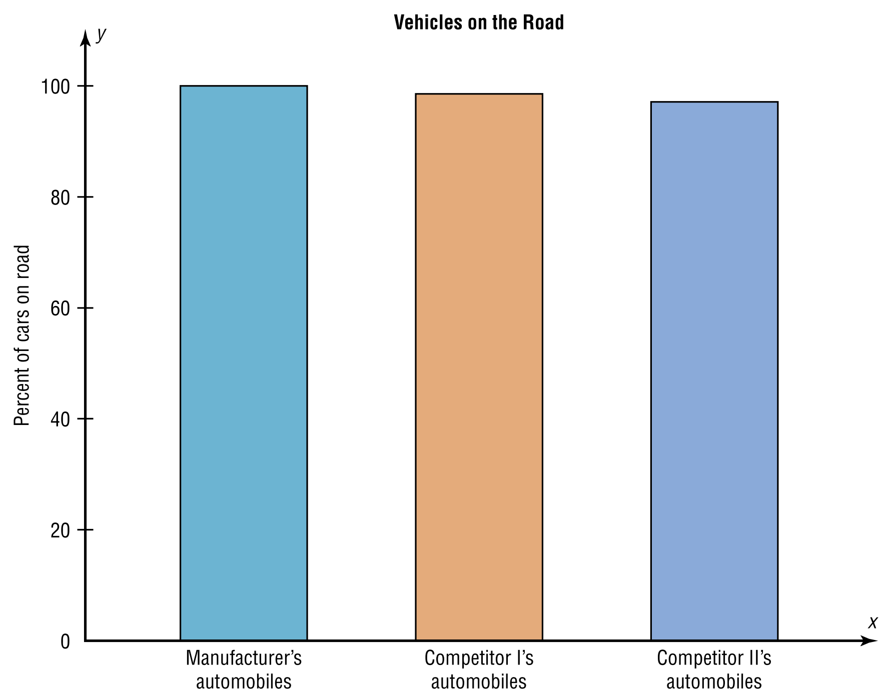


Figure 2–17

Graph in Figure 2–16 Redrawn Using a Scale from 0 to 100%



It is not wrong to truncate an axis of the graph; many times it is necessary to do so. However, the reader should be aware of this fact and interpret the graph accordingly. Do not be misled if an inappropriate impression is given.

Let us consider another example. The projected required fuel economy in miles per gallon for General Motors vehicles is shown. In this case, an increase from 21.9 to 23.2 miles per gallon is projected.

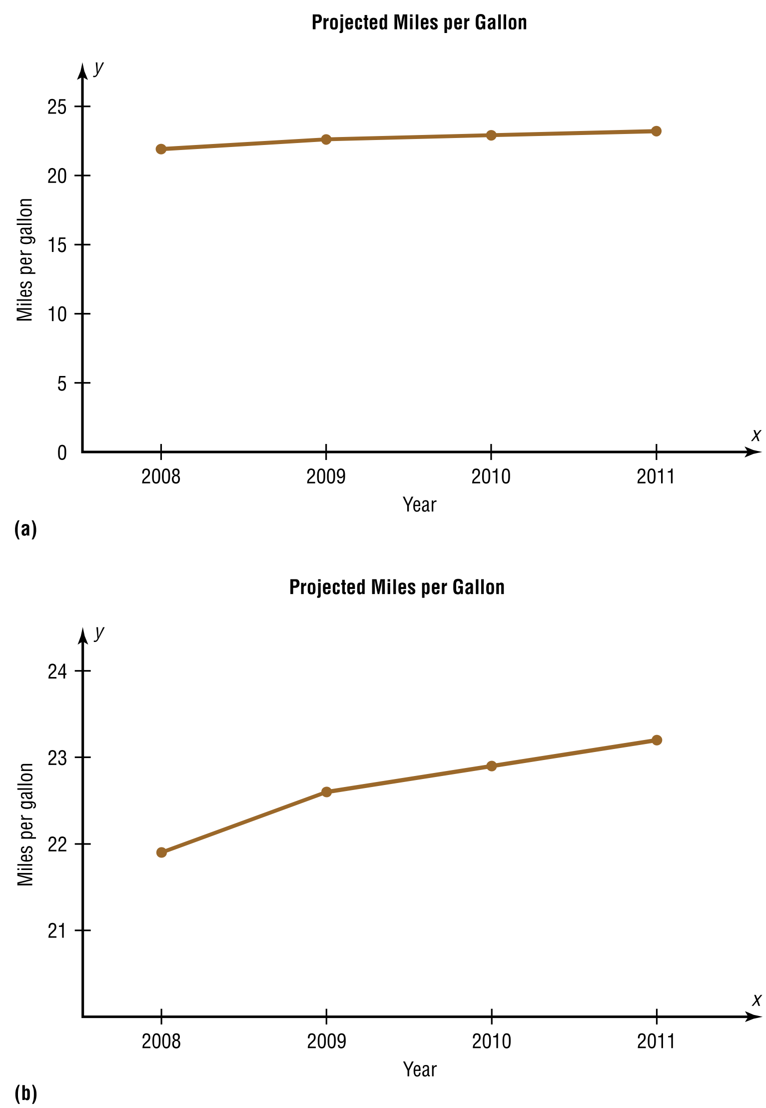
Year	2008	2009	2010	2011
MPG	21.9	22.6	22.9	23.2

Source: National Highway Traffic Safety Administration.

When you examine the graph shown in Figure 2–18(a) using a scale of 0 to 25 miles per gallon, the graph shows a slight increase. However, when the scale is changed to 21

Figure 2–18

Projected Miles per Gallon



to 24 miles per gallon, the graph shows a much larger increase even though the data remain the same. See Figure 2–18(b). Again, by changing the units or starting point on the y axis, one can change the visual representation.

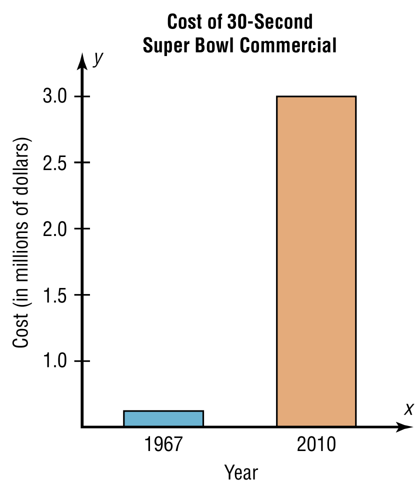
Another misleading graphing technique sometimes used involves exaggerating a one-dimensional increase by showing it in two dimensions. For example, the average cost of a 30-second Super Bowl commercial has increased from $42,000 in 1967 to $3 million in 2010 (Source: *USA TODAY*).

The increase shown by the graph in Figure 2–19(a) represents the change by a comparison of the heights of the two bars in one dimension. The same data are shown two-dimensionally with circles in Figure 2–19(b). Notice that the difference seems much larger because the eye is comparing the areas of the circles rather than the lengths of the diameters.

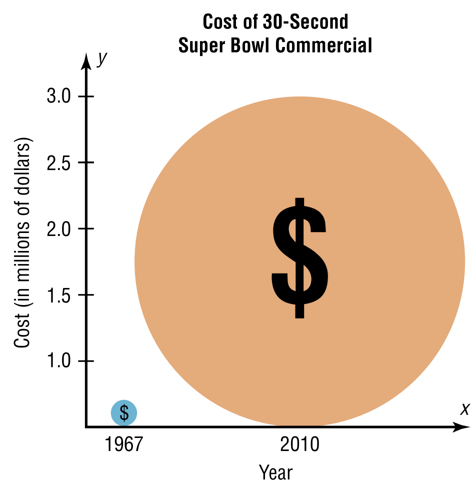
Note that it is not wrong to use the graphing techniques of truncating the scales or representing data by two-dimensional pictures. But when these techniques are used, the reader should be cautious of the conclusion drawn on the basis of the graphs.

Figure 2–19

Comparison of Costs for a 30-Second Super Bowl Commercial



(a) Graph using bars

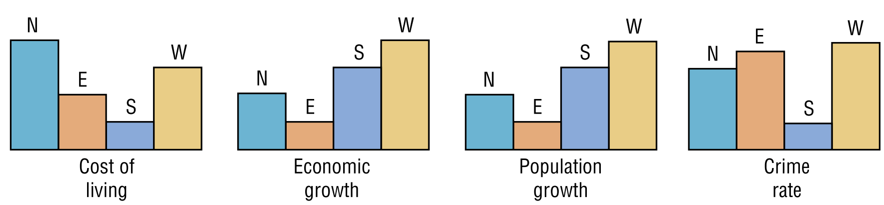


(b) Graph using circles

Another way to misrepresent data on a graph is by omitting labels or units on the axes of the graph. The graph shown in Figure 2–20 compares the cost of living, economic growth, population growth, etc., of four main geographic areas in the United States. However, since there are no numbers on the y axis, very little information can be gained from this graph, except a crude ranking of each factor. There is no way to decide the actual magnitude of the differences.

Figure 2–20

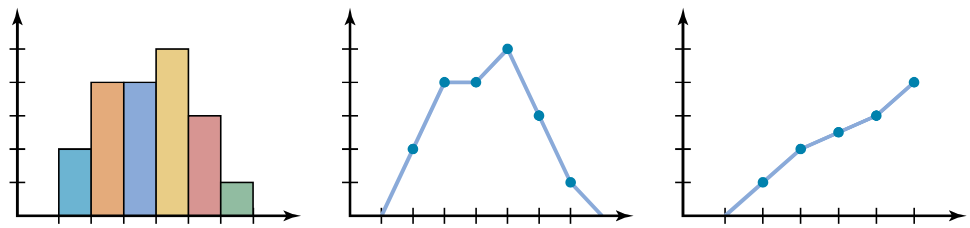
A Graph with No Units on the *y* Axis



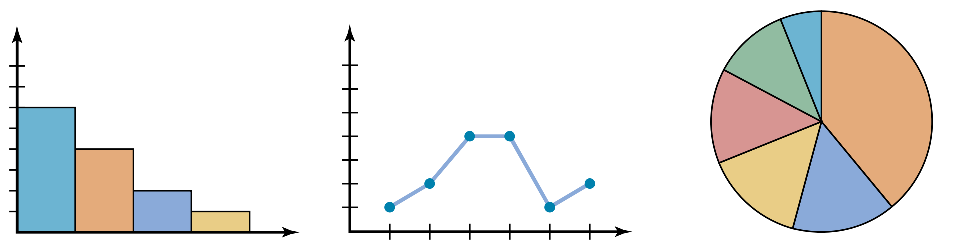
Finally, all graphs should contain a source for the information presented. The inclusion of a source for the data will enable you to check the reliability of the organization presenting the data. A summary of the types of graphs and their uses is shown in Figure 2–21.

Figure 2–21

Summary of Graphs and Uses of Each



(a) Histogram; frequency polygon; ogive
Used when the data are contained in a grouped frequency distribution.



(b) Pareto chart
Used to show frequencies for nominal or qualitative variables.

(c) Time series graph
Used to show a pattern or trend that occurs over a period of time.

(d) Pie graph
Used to show the relationship between the parts and the whole. (Most often uses percentages.)

Stem and Leaf Plots

The stem and leaf plot is a method of organizing data and is a combination of sorting and graphing. It has the advantage over a grouped frequency distribution of retaining the actual data while showing them in graphical form.

Objective 4

Draw and interpret a stem and leaf plot.

A **stem and leaf plot** is a data plot that uses part of the data value as the stem and part of the data value as the leaf to form groups or classes.

Example 2–13 shows the procedure for constructing a stem and leaf plot.

Example 2–13

At an outpatient testing center, the number of cardiograms performed each day for 20 days is shown. Construct a stem and leaf plot for the data.

25	31	20	32	13
14	43	02	57	23
36	32	33	32	44
32	52	44	51	45

Speaking of
Statistics

How Much Paper Money Is in Circulation Today?

The Federal Reserve estimated that during a recent year, there were 22 billion bills in circulation. About 35% of them were \$1 bills, 3% were \$2 bills, 8% were \$5 bills, 7% were \$10 bills, 23% were \$20 bills, 5% were \$50 bills, and 19% were \$100 bills. It costs about 3¢ to print each \$1 bill.

The average life of a \$1 bill is 22 months, a \$10 bill 3 years, a \$20 bill 4 years, a \$50 bill 9 years, and a \$100 bill 9 years. What type of graph would you use to represent the average lifetimes of the bills?

Solution

Step 1 Arrange the data in order:

02, 13, 14, 20, 23, 25, 31, 32, 32, 32,
32, 33, 36, 43, 44, 44, 45, 51, 52, 57

Note: Arranging the data in order is not essential and can be cumbersome when the data set is large; however, it is helpful in constructing a stem and leaf plot. The leaves in the final stem and leaf plot should be arranged in order.

Step 2 Separate the data according to the first digit, as shown.

02 13, 14 20, 23, 25 31, 32, 32, 32, 32, 33, 36
43, 44, 44, 45 51, 52, 57

Step 3 A display can be made by using the leading digit as the *stem* and the trailing digit as the *leaf.* For example, for the value 32, the leading digit, 3, is the stem and the trailing digit, 2, is the leaf. For the value 14, the 1 is the stem and the 4 is the leaf. Now a plot can be constructed as shown in Figure 2–22.

Leading digit (stem)	Trailing digit (leaf)
0	2
1	3 4
2	0 3 5
3	1 2 2 2 2 3 6
4	3 4 4 5
5	1 2 7

Figure 2–22
Stem and Leaf Plot for Example 2–13

0	2
1	3 4
2	0 3 5
3	1 2 2 2 2 3 6
4	3 4 4 5
5	1 2 7

Figure 2–22 shows that the distribution peaks in the center and that there are no gaps in the data. For 7 of the 20 days, the number of patients receiving cardiograms was between 31 and 36. The plot also shows that the testing center treated from a minimum of 2 patients to a maximum of 57 patients in any one day.

If there are no data values in a class, you should write the stem number and leave the leaf row blank. Do not put a zero in the leaf row.

Example 2–14

An insurance company researcher conducted a survey on the number of car thefts in a large city for a period of 30 days last summer. The raw data are shown. Construct a stem and leaf plot by using classes 50–54, 55–59, 60–64, 65–69, 70–74, and 75–79.

52	62	51	50	69
58	77	66	53	57
75	56	55	67	73
79	59	68	65	72
57	51	63	69	75
65	53	78	66	55

Solution

Step 1 Arrange the data in order.

50, 51, 51, 52, 53, 53, 55, 55, 56, 57, 57, 58, 59, 62, 63, 65, 65, 66, 66, 67, 68, 69, 69, 72, 73, 75, 75, 77, 78, 79

Step 2 Separate the data according to the classes.

50, 51, 51, 52, 53, 53   55, 55, 56, 57, 57, 58, 59
62, 63   65, 65, 66, 66, 67, 68, 69, 69   72, 73
75, 75, 77, 78, 79

Step 3 Plot the data as shown here.

Leading digit (stem)	Trailing digit (leaf)
5	0 1 1 2 3 3
5	5 5 6 7 7 8 9
6	2 3
6	5 5 6 6 7 8 9 9
7	2 3
7	5 5 7 8 9

The graph for this plot is shown in Figure 2–23.

Figure 2–23

Stem and Leaf Plot for Example 2–14

5	0 1 1 2 3 3
5	5 5 6 7 7 8 9
6	2 3
6	5 5 6 6 7 8 9 9
7	2 3
7	5 5 7 8 9

When the data values are in the hundreds, such as 325, the stem is 32 and the leaf is 5. For example, the stem and leaf plot for the data values 325, 327, 330, 332, 335, 341, 345, and 347 looks like this.

32	5 7
33	0 2 5
34	1 5 7

When you analyze a stem and leaf plot, look for peaks and gaps in the distribution. See if the distribution is symmetric or skewed. Check the variability of the data by looking at the spread.

Interesting Fact

The average number of pencils and index cards David Letterman tosses over his shoulder during one show is 4.

Related distributions can be compared by using a back-to-back stem and leaf plot. The back-to-back stem and leaf plot uses the same digits for the stems of both distributions, but the digits that are used for the leaves are arranged in order out from the stems on both sides. Example 2–15 shows a back-to-back stem and leaf plot.

Example 2–15

The number of stories in two selected samples of tall buildings in Atlanta and Philadelphia is shown. Construct a back-to-back stem and leaf plot, and compare the distributions.

Atlanta					Philadelphia				
55	70	44	36	40	61	40	38	32	30
63	40	44	34	38	58	40	40	25	30
60	47	52	32	32	54	40	36	30	30
50	53	32	28	31	53	39	36	34	33
52	32	34	32	50	50	38	36	39	32
26	29								

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

Solution

Step 1 Arrange the data for both data sets in order.

Step 2 Construct a stem and leaf plot using the same digits as stems. Place the digits for the leaves for Atlanta on the left side of the stem and the digits for the leaves for Philadelphia on the right side, as shown. See Figure 2–24.

Figure 2–24

Back-to-Back Stem and Leaf Plot for Example 2–15

Atlanta		Philadelphia
9 8 6	2	5
8 6 4 4 2 2 2 2 2 1	3	0 0 0 0 2 2 3 4 6 6 6 8 8 9 9
7 4 4 0 0	4	0 0 0 0
5 3 2 2 0 0	5	0 3 4 8
3 0	6	1
0	7	

Step 3 Compare the distributions. The buildings in Atlanta have a large variation in the number of stories per building. Although both distributions are peaked in the 30- to 39-story class, Philadelphia has more buildings in this class. Atlanta has more buildings that have 40 or more stories than Philadelphia does.

Stem and leaf plots are part of the techniques called *exploratory data analysis.* More information on this topic is presented in Chapter 3.

Applying the Concepts 2–3

Leading Cause of Death

The following shows approximations of the leading causes of death among men ages 25–44 years. The rates are per 100,000 men. Answer the following questions about the graph.