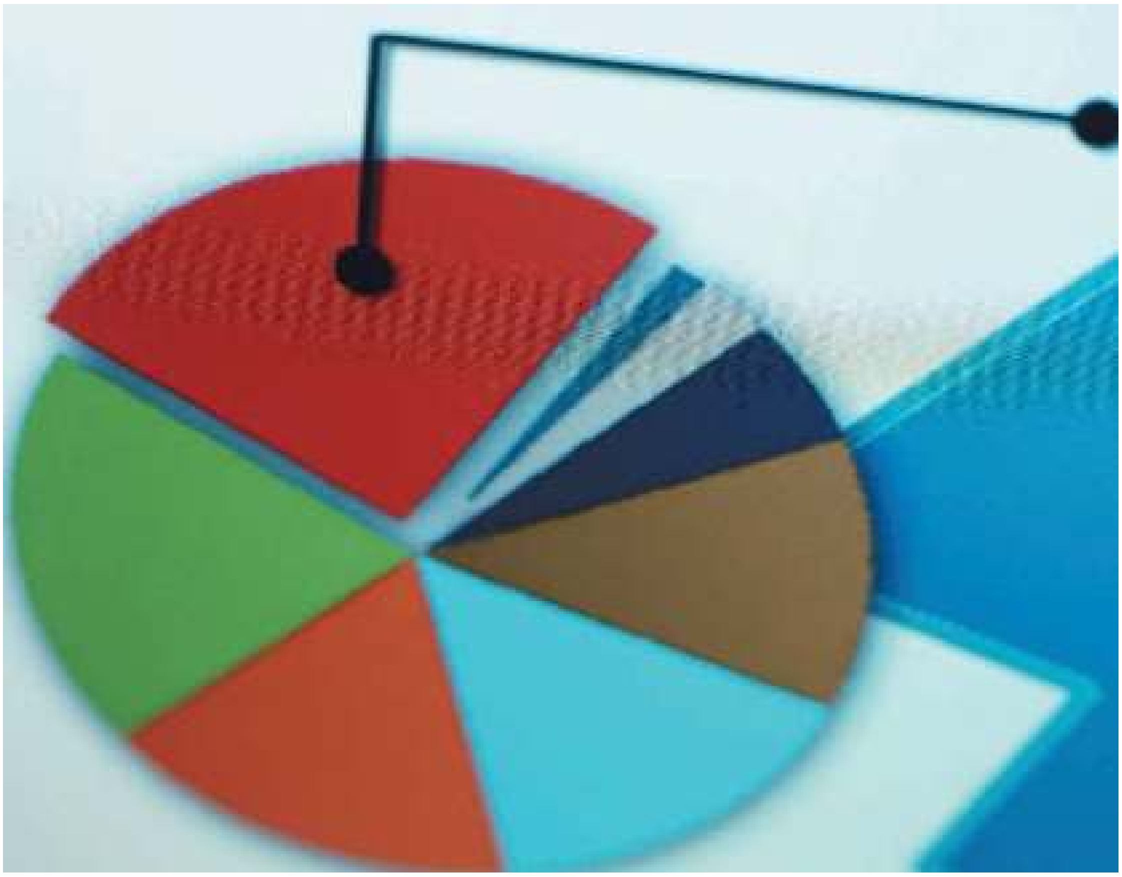
CHAPTER 2

Charts and Graphs

LEARNING OBJECTIVES

The overall objective of Chapter 2 is for you to master several techniques for summarizing and depicting data, thereby enabling you to:

1. Construct a frequency distribution from a set of data
2. Construct different types of quantitative data graphs, including histograms, frequency polygons, ogives, dot plots, and stem-and-leaf plots, in order to interpret the data being graphed
3. Construct different types of qualitative data graphs, including pie charts, bar graphs, and Pareto charts, in order to interpret the data being graphed
4. Recognize basic trends in two-variable scatter plots of numerical data



Photodisc/SUPERSTOCK

Energy Consumption Around the World

As most people suspect, the United States is the number one consumer of oil in the world, followed by China, Japan, India, Russia, Germany, South Korea, and Canada. China, however, is the world's largest consumer of coal, with the United States coming in number two, followed by India, Japan, and Russia. The annual oil and coal consumption figures for eight of the top total energy-consuming nations in the world, according to figures released by the *BP Statistical Review of World Energy* for a recent year, are as follows.

Country	Oil Consumption (million tons)	Coal Consumption (million tons oil equivalent)
United States	943.1	573.7
China	368.0	1311.4
Japan	228.9	125.3
India	128.5	208.0
Russia	125.9	94.5
Germany	112.5	86.0
South Korea	107.6	59.7
Canada	102.3	30.4

(*continued*)

Managerial and Statistical Questions

Suppose you are an energy industry analyst and you are asked to prepare a brief report showing the leading energy-consumption countries in both oil and coal.

1. What is the best way to display the energy consumption data in a report? Are the raw data enough? Can you effectively display the data graphically?
2. Is there a way to graphically display oil and coal figures together so that readers can visually compare countries on their consumptions of the two different energy sources?

Source: BP Statistical Review of World Energy, June 2008, http://www.bp.com/liveassets/bp_internet/globalbp/globalbp_uk_english/reports_and_publications/statistical_energy_review_2008/STAGING/local_assets/downloads/pdf/statistical_review_of_world_energy_full_review_2008.pdf

In Chapters 2 and 3 many techniques are presented for reformatting or reducing data so that the data are more manageable and can be used to assist decision makers more effectively. Two techniques for grouping data are the frequency distribution and the stem-and-leaf plot presented in this chapter. In addition, Chapter 2 discusses and displays several graphical tools for summarizing and presenting data, including histogram, frequency polygon, ogive, dot plot, bar chart, pie chart, and Pareto chart for one-variable data and the scatter plot for two-variable numerical data.

Raw data, or *data that have not been summarized in any way,* are sometimes referred to as **ungrouped data.** Table 2.1 contains 60 years of raw data of the unemployment rates for Canada. *Data that have been organized into a frequency distribution* are called **grouped data.** Table 2.2 presents a frequency distribution for the data displayed in Table 2.1. The distinction

TABLE 2.1

60 Years of Canadian Unemployment Rates (ungrouped data)

2.3	7.0	6.3	11.3	9.6
2.8	7.1	5.6	10.6	9.1
3.6	5.9	5.4	9.7	8.3
2.4	5.5	7.1	8.8	7.6
2.9	4.7	7.1	7.8	6.8
3.0	3.9	8.0	7.5	7.2
4.6	3.6	8.4	8.1	7.7
4.4	4.1	7.5	10.3	7.6
3.4	4.8	7.5	11.2	7.2
4.6	4.7	7.6	11.4	6.8
6.9	5.9	11.0	10.4	6.3
6.0	6.4	12.0	9.5	6.0

TABLE 2.2

Frequency Distribution of 60 Years of Unemployment Data for Canada (grouped data)

Class Interval	Frequency
1–under 3	4
3–under 5	12
5–under 7	13
7–under 9	19
9–under 11	7
11–under 13	5

between ungrouped and grouped data is important because the calculation of statistics differs between the two types of data. This chapter focuses on organizing ungrouped data into grouped data and displaying them graphically.

2.1 FREQUENCY DISTRIBUTIONS

One particularly useful tool for grouping data is the **frequency distribution,** which is *a summary of data presented in the form of class intervals and frequencies.* How is a frequency distribution constructed from raw data? That is, how are frequency distributions like the one displayed in Table 2.2 constructed from raw data like those presented in Table 2.1? Frequency distributions are relatively easy to construct. Although some guidelines and rules of thumb help in their construction, frequency distributions vary in final shape and design, even when the original raw data are identical. In a sense, frequency distributions are constructed according to individual business researchers' taste.

When constructing a frequency distribution, the business researcher should first determine the range of the raw data. The **range** often is defined as *the difference between the largest and smallest numbers.* The range for the data in Table 2.1 is 9.7 (12.0–2.3).

The second step in constructing a frequency distribution is to determine how many classes it will contain. One rule of thumb is to select between *5 and 15 classes*. If the frequency distribution contains too few classes, the data summary may be too general to be useful. Too many classes may result in a frequency distribution that does not aggregate the data enough to be helpful. The final number of classes is arbitrary. The business researcher arrives at a number by examining the range and determining a number of classes that will span the range adequately and also be meaningful to the user. The data in Table 2.1 were grouped into six classes for Table 2.2.

After selecting the number of classes, the business researcher must determine the width of the class interval. An approximation of the class width can be calculated by dividing the range by the number of classes. For the data in Table 2.1, this approximation would be $9.7/6 = 1.62$. Normally, the number is rounded up to the next whole number, which in this case is 2. The frequency distribution must start at a value equal to or lower than the lowest number of the ungrouped data and end at a value equal to or higher than the highest number. The lowest unemployment rate is 2.3 and the highest is 12.0, so the business researcher starts the frequency distribution at 1 and ends it at 13. Table 2.2 contains the completed frequency distribution for the data in Table 2.1. Class endpoints are selected so that no value of the data can fit into more than one class. The class interval expression "under" in the distribution of Table 2.2 avoids such a problem.

Class Midpoint

The *midpoint of each class interval* is called the **class midpoint** and is sometimes referred to as the **class mark**. It is *the value halfway across the class interval* and can be calculated as *the average of the two class endpoints*. For example, in the distribution of Table 2.2, the midpoint of the class interval 3–under 5 is 4, or $(3 + 5)/2$.

The class midpoint is important, because it becomes the representative value for each class in most group statistics calculations. The third column in Table 2.3 contains the class midpoints for all classes of the data from Table 2.2.

Relative Frequency

Relative frequency is *the proportion of the total frequency that is in any given class interval in a frequency distribution.* Relative frequency is the individual class frequency divided by the total frequency. For example, from Table 2.3, the relative frequency for the class interval 5–under 7 is $13/60 = .2167$. Consideration of the relative frequency is preparatory to the study of probability in Chapter 4. Indeed, if values were selected randomly from the data in Table 2.1, the probability of drawing a number that is "5–under 7" would be .2167, the

TABLE 2.3

Class Midpoints, Relative Frequencies, and Cumulative Frequencies for Unemployment Data

Interval	Frequency	Class Midpoint	Relative Frequency	Cumulative Frequency
1–under 3	4	2	.0667	4
3–under 5	12	4	.2000	16
5–under 7	13	6	.2167	29
7–under 9	19	8	.3167	48
9–under 11	7	10	.1167	55
11–under 13	5	12	.0833	60
Total	60			

relative frequency for that class interval. The fourth column of Table 2.3 lists the relative frequencies for the frequency distribution of Table 2.2.

Cumulative Frequency

The **cumulative frequency** is *a running total of frequencies through the classes of a frequency distribution*. The cumulative frequency for each class interval is the frequency for that class interval added to the preceding cumulative total. In Table 2.3 the cumulative frequency for the first class is the same as the class frequency: 4. The cumulative frequency for the second class interval is the frequency of that interval (12) plus the frequency of the first interval (4), which yields a new cumulative frequency of 16. This process continues through the last interval, at which point the cumulative total equals the sum of the frequencies (60). The concept of cumulative frequency is used in many areas, including sales cumulated over a fiscal year, sports scores during a contest (cumulated points), years of service, points earned in a course, and costs of doing business over a period of time. Table 2.3 gives cumulative frequencies for the data in Table 2.2.

DEMONSTRATION PROBLEM 2.1

Demonstration Problem

The following data are the average weekly mortgage interest rates for a 40-week period.

7.29	7.23	7.11	6.78	7.47
6.69	6.77	6.57	6.80	6.88
6.98	7.16	7.30	7.24	7.16
7.03	6.90	7.16	7.40	7.05
7.28	7.31	6.87	7.68	7.03
7.17	6.78	7.08	7.12	7.31
7.40	6.35	6.96	7.29	7.16
6.97	6.96	7.02	7.13	6.84

Construct a frequency distribution for these data. Calculate and display the class midpoints, relative frequencies, and cumulative frequencies for this frequency distribution.

Solution

How many classes should this frequency distribution contain? The range of the data is 1.33 (7.68–6.35). If 7 classes are used, each class width is approximately:

$$\text{Class Width} = \frac{\text{Range}}{\text{Number of Classes}} = \frac{1.33}{7} = 0.19$$

If a class width of .20 is used, a frequency distribution can be constructed with endpoints that are more uniform looking and allow presentation of the information in categories more familiar to mortgage interest rate users.

The first class endpoint must be 6.35 or lower to include the smallest value; the last endpoint must be 7.68 or higher to include the largest value. In this case the frequency distribution begins at 6.30 and ends at 7.70. The resulting frequency distribution, class midpoints, relative frequencies, and cumulative frequencies are listed in the following table.

Interval	Frequency	Class Midpoint	Relative Frequency	Cumulative Frequency
6.30–under 6.50	1	6.40	.025	1
6.50–under 6.70	2	6.60	.050	3
6.70–under 6.90	7	6.80	.175	10
6.90–under 7.10	10	7.00	.250	20
7.10–under 7.30	13	7.20	.325	33
7.30–under 7.50	6	7.40	.150	39
7.50–under 7.70	1	7.60	.025	40
Total	40			

The frequencies and relative frequencies of these data reveal the mortgage interest rate classes that are likely to occur during the period. Most of the mortgage interest rates (36 of the 40) are in the classes starting with 6.70–under 6.90 and going through 7.30–under 7.50. The rates with the greatest frequency, 13, are in the 7.10-under 7.30 class.

2.1 PROBLEMS

2.1 The following data represent the afternoon high temperatures for 50 construction days during a year in St. Louis.

42	70	64	47	66	69	73	38	48	25
55	85	10	24	45	31	62	47	63	84
16	40	81	15	35	17	40	36	44	17
38	79	35	36	23	64	75	53	31	60
31	38	52	16	81	12	61	43	30	33

a. Construct a frequency distribution for the data using five class intervals.

b. Construct a frequency distribution for the data using 10 class intervals.

c. Examine the results of (a) and (b) and comment on the usefulness of the frequency distribution in terms of temperature summarization capability.

2.2 A packaging process is supposed to fill small boxes of raisins with approximately 50 raisins so that each box will weigh the same. However, the number of raisins in each box will vary. Suppose 100 boxes of raisins are randomly sampled, the raisins counted, and the following data are obtained.

57	51	53	52	50	60	51	51	52	52
44	53	45	57	39	53	58	47	51	48
49	49	44	54	46	52	55	54	47	53
49	52	49	54	57	52	52	53	49	47
51	48	55	53	55	47	53	43	48	46
54	46	51	48	53	56	48	47	49	57
55	53	50	47	57	49	43	58	52	44
46	59	57	47	61	60	49	53	41	48
59	53	45	45	56	40	46	49	50	57
47	52	48	50	45	56	47	47	48	46

Construct a frequency distribution for these data. What does the frequency distribution reveal about the box fills?

2.3 The owner of a fast-food restaurant ascertains the ages of a sample of customers. From these data, the owner constructs the frequency distribution shown. For each class interval of the frequency distribution, determine the class midpoint, the relative frequency, and the cumulative frequency.

Class Interval	Frequency
0–under 5	6
5–under 10	8
10–under 15	17
15–under 20	23
20–under 25	18
25–under 30	10
30–under 35	4

What does the relative frequency tell the fast-food restaurant owner about customer ages?

2.4 The human resources manager for a large company commissions a study in which the employment records of 500 company employees are examined for absenteeism during the past year. The business researcher conducting the study organizes the data into a frequency distribution to assist the human resources manager in analyzing the data. The frequency distribution is shown. For each class of the frequency distribution, determine the class midpoint, the relative frequency, and the cumulative frequency.

Class Interval	Frequency
0–under 2	218
2–under 4	207
4–under 6	56
6–under 8	11
8–under 10	8

2.5 List three specific uses of cumulative frequencies in business.

2.2 QUANTITATIVE DATA GRAPHS

Interactive Applet

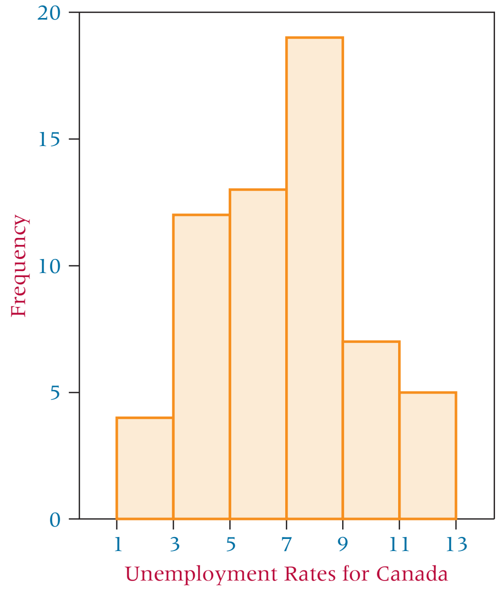
One of the most effective mechanisms for presenting data in a form meaningful to decision makers is graphical depiction. Through graphs and charts, the decision maker can often get an overall picture of the data and reach some useful conclusions merely by studying the chart or graph. Converting data to graphics can be creative and artful. Often the most difficult step in this process is to reduce important and sometimes expensive data to a graphic picture that is both clear and concise and yet consistent with the message of the original data. One of the most important uses of graphical depiction in statistics is to help the researcher determine the shape of a distribution. Data graphs can generally be classified as quantitative or qualitative. Quantitative data graphs are plotted along a numerical scale, and qualitative graphs are plotted using non-numerical categories. In this section, we will examine five types of quantitative data graphs: (1) histogram, (2) frequency polygon, (3) ogive, (4) dot plot, and (5) stem-and-leaf plot.

Histograms

One of the more widely used types of graphs for quantitative data is the **histogram**. A histogram is a series of contiguous bars or rectangles that represent the frequency of data in given class intervals. If the class intervals used along the horizontal axis are equal, then the height of the bars represent the frequency of values in a given class interval. If the class intervals are unequal, then the areas of the bars (rectangles) can be used for relative comparisons of class frequencies. Construction of a histogram involves labeling the x-axis (abscissa) with the class endpoints and the y-axis (ordinate) with the frequencies, drawing a horizontal line segment from class endpoint to class endpoint at each frequency value, and connecting each line segment vertically from the frequency value to the x-axis to form a series of rectangles (bars). Figure 2.1 is a histogram of the frequency distribution in Table 2.2 produced by using the software package Minitab.

FIGURE 2.1

Minitab Histogram of Canadian Unemployment Data



A histogram is a useful tool for differentiating the frequencies of class intervals. A quick glance at a histogram reveals which class intervals produce the highest frequency totals. Figure 2.1 clearly shows that the class interval 7–under 9 yields by far the highest frequency count (19). Examination of the histogram reveals where large increases or decreases occur between classes, such as from the 1–under 3 class to the 3–under 5 class, an increase of 8, and from the 7–under 9 class to the 9–under 11 class, a decrease of 12.

Note that the scales used along the *x*-and *y*-axes for the histogram in Figure 2.1 are almost identical. However, because ranges of meaningful numbers for the two variables being graphed often differ considerably, the graph may have different scales on the two axes. Figure 2.2 shows what the histogram of unemployment rates would look like if the scale on the *y*-axis were more compressed than that on the *x*-axis. Notice that less difference in the length of the rectangles appears to represent the frequencies in Figure 2.2. It is important that the user of the graph clearly understands the scales used for the axes of a histogram. Otherwise, a graph's creator can "lie with statistics" by stretching or compressing a graph to make a point.*

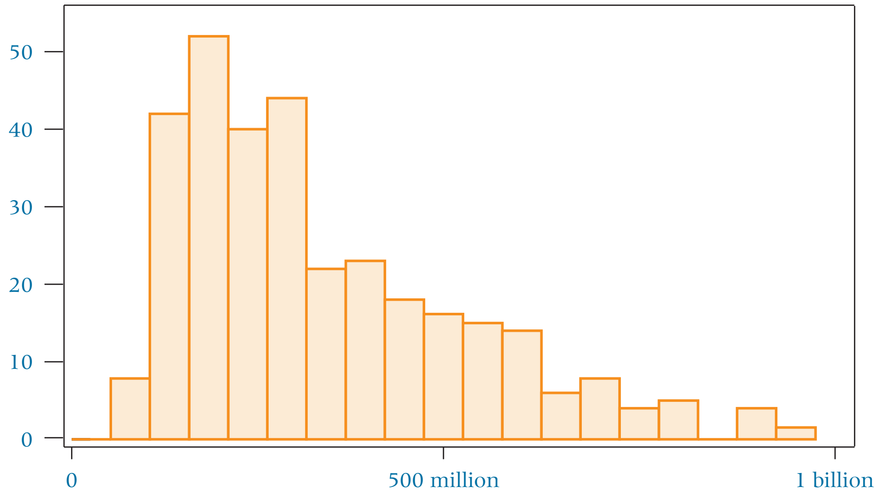
FIGURE 2.2

Minitab Histogram of Canadian Unemployment Data (*y*-axis compressed)

*It should be pointed out that the software package Excel uses the term *histogram* to refer to a frequency distribution. However, by checking Chart Output in the Excel histogram dialog box, a graphical histogram is also created.

FIGURE 2.3

Histogram of Stock Volumes



Using Histograms to Get an Initial Overview of the Data

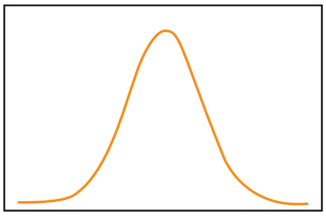
Because of the widespread availability of computers and statistical software packages to business researchers and decision makers, the histogram continues to grow in importance in yielding information about the shape of the distribution of a large database, the variability of the data, the central location of the data, and outlier data. Although most of these concepts are presented in Chapter 3, the notion of histogram as an initial tool to access these data characteristics is presented here.

A business researcher measured the volume of stocks traded on Wall Street three times a month for nine years resulting in a database of 324 observations. Suppose a financial decision maker wants to use these data to reach some conclusions about the stock market. Figure 2.3 shows a Minitab-produced histogram of these data. What can we learn from this histogram? Virtually all stock market volumes fall between zero and 1 billion shares. The distribution takes on a shape that is high on the left end and tapered to the right. In Chapter 3 we will learn that the shape of this distribution is skewed toward the right end. In statistics, it is often useful to determine whether data are approximately normally distributed (bell-shaped curve) as shown in Figure 2.4. We can see by examining the histogram in Figure 2.3 that the stock market volume data are not normally distributed. Although the center of the histogram is located near 500 million shares, a large portion of stock volume observations falls in the lower end of the data somewhere between 100 million and 400 million shares. In addition, the histogram shows some outliers in the upper end of the distribution. Outliers are data points that appear outside of the main body of observations and may represent phenomena that differ from those represented by other data points. By observing the histogram, we notice a few data observations near 1 billion. One could conclude that on a few stock market days an unusually large volume of shares are traded. These and other insights can be gleaned by examining the histogram and show that histograms play an important role in the initial analysis of data.

Frequency Polygons

FIGURE 2.4

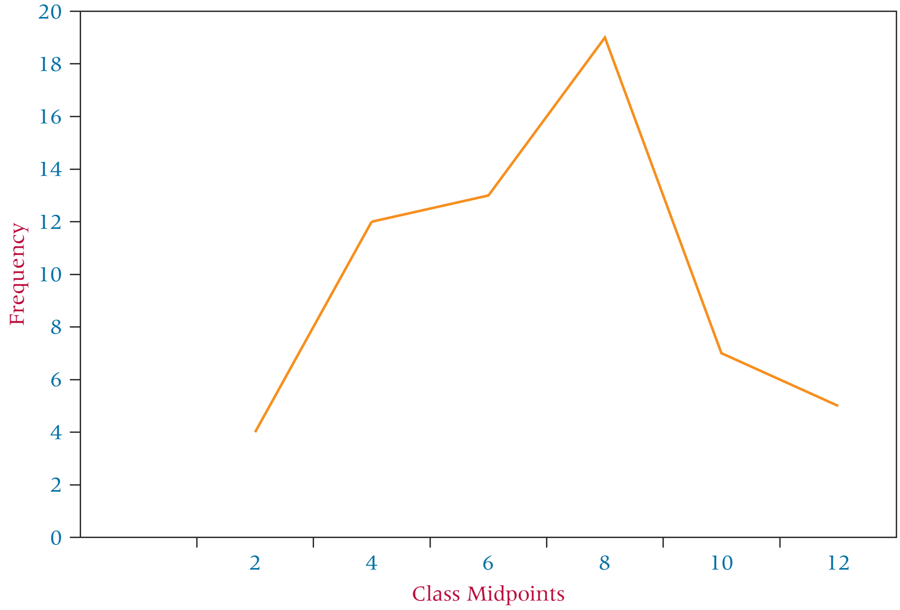
Normal Distribution



A **frequency polygon**, like the histogram, is a graphical display of class frequencies. However, instead of using bars or rectangles like a histogram, in a frequency polygon each class frequency is plotted as a dot at the class midpoint, and the dots are connected by a series of line segments. Construction of a frequency polygon begins by scaling class midpoints along the horizontal axis and the frequency scale along the vertical axis. A dot is plotted for the associated frequency value at each class midpoint. Connecting these midpoint dots completes the graph. Figure 2.5 shows a frequency polygon of the distribution data from Table 2.2 produced by using the software package Excel. The information gleaned from frequency polygons and histograms is similar. As with the histogram, changing the scales of the axes can compress or stretch a frequency polygon, which affects the user's impression of what the graph represents.

FIGURE 2.5

Excel-Produced Frequency Polygon of the Unemployment Data



Ogives

An **ogive** (o-jive) is *a cumulative frequency polygon*. Construction begins by labeling the *x*-axis with the class endpoints and the *y*-axis with the frequencies. However, the use of cumulative frequency values requires that the scale along the *y*-axis be great enough to include the frequency total. A dot of zero frequency is plotted at the beginning of the first class, and construction proceeds by marking a dot at the *end* of each class interval for the cumulative value. Connecting the dots then completes the ogive. Figure 2.6 presents an ogive produced by using Excel for the data in Table 2.2.

Ogives are most useful when the decision maker wants to see *running totals*. For example, if a comptroller is interested in controlling costs, an ogive could depict cumulative costs over a fiscal year.

Steep slopes in an ogive can be used to identify sharp increases in frequencies. In Figure 2.6, a particularly steep slope occurs in the 7–under 9 class, signifying a large jump in class frequency totals.

FIGURE 2.6

Excel Ogive of the Unemployment Data

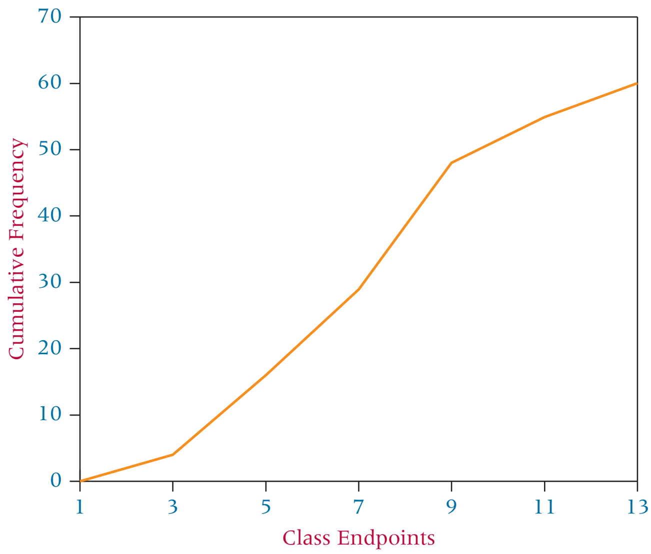


FIGURE 2.7

A Minitab-Produced Dot Plot of the Canadian Unemployment Data

Dot Plots

A relatively simple statistical chart that is generally used to display continuous, quantitative data is the **dot plot**. In a dot plot, each data value is plotted along the horizontal axis and is represented on the chart by a dot. If multiple data points have the same values, the dots will stack up vertically. If there are a large number of close points, it may not be possible to display all of the data values along the horizontal axis. Dot plots can be especially useful for observing the overall shape of the distribution of data points along with identifying data values or intervals for which there are groupings and gaps in the data. Figure 2.7 displays a minitab-produced dot plot for the Canadian unemployment data shown in Table 2.1. Note that the distribution is relatively balanced with a peak near the center. There are a few gaps to note, such as from 4.9 to 5.3, from 9.9 to 10.2, and from 11.5 to 11.9. In addition, there are groupings around 6.0, 7.1, and 7.5.

Stem-and-Leaf Plots

Another way to organize raw data into groups besides using a frequency distribution is a **stem-and-leaf plot**. This technique is simple and provides a unique view of the data. A stem-and-leaf plot is constructed by separating the digits for each number of the data into two groups, *a stem and a leaf.* The leftmost digits are the stem and consist of the higher valued digits. The rightmost digits are the leaves and contain the lower values. If a set of data has only two digits, the stem is the value on the left and the leaf is the value on the right. For example, if 34 is one of the numbers, the stem is 3 and the leaf is 4. For numbers with more than two digits, division of stem and leaf is a matter of researcher preference.

Table 2.4 contains scores from an examination on plant safety policy and rules given to a group of 35 job trainees. A stem-and-leaf plot of these data is displayed in Table 2.5. One advantage of such a distribution is that the instructor can readily see whether the scores are in the upper or lower end of each bracket and also determine the spread of the scores. A second advantage of stem-and-leaf plots is that the values of the original raw data are retained (whereas most frequency distributions and graphic depictions use the class midpoint to represent the values in a class).

TABLE 2.4

Safety Examination Scores for Plant Trainees

86	77	91	60	55
76	92	47	88	67
23	59	72	75	83
77	68	82	97	89
81	75	74	39	67
79	83	70	78	91
68	49	56	94	81

TABLE 2.5

Stem and Leaf Plot for Plant Safety Examination Data

Stem	Leaf									
2	3									
3	9									
4	7	9								
5	5	6	9							
6	0	7	7	8	8					
7	0	2	4	5	5	6	7	7	8	9
8	1	1	2	3	3	6	8	9		
9	1	1	2	4	7					

DEMONSTRATION PROBLEM 2.2

The following data represent the costs (in dollars) of a sample of 30 postal mailings by a company.

3.67	2.75	9.15	5.11	3.32	2.09
1.83	10.94	1.93	3.89	7.20	2.78
6.72	7.80	5.47	4.15	3.55	3.53
3.34	4.95	5.42	8.64	4.84	4.10
5.10	6.45	4.65	1.97	2.84	3.21

Using dollars as a stem and cents as a leaf, construct a stem-and-leaf plot of the data.

Solution

Stem	Leaf						
1	83	93	97				
2	09	75	78	84			
3	21	32	34	53	55	67	89
4	10	15	65	84	95		
5	10	11	42	47			
6	45	72					
7	20	80					
8	64						
9	15						
10	94						

2.2 PROBLEMS

2.6 Construct a histogram and a frequency polygon for the following data.

Class Interval	Frequency
30–under 32	5
32–under 34	7
34–under 36	15
36–under 38	21
38–under 40	34
40–under 42	24
42–under 44	17
44–under 46	8

2.7 Construct a histogram and a frequency polygon for the following data.

Class Interval	Frequency
10–under 20	9
20–under 30	7
30–under 40	10
40–under 50	6
50–under 60	13
60–under 70	18
70–under 80	15

2.8 Construct an ogive for the following data.

Class Interval	Frequency
3–under 6	2
6–under 9	5
9–under 12	10
12–under 15	11
15–under 18	17
18–under 21	5

2.9 Construct a stem-and-leaf plot using two digits for the stem.

212	239	240	218	222	249	265	224
257	271	266	234	239	219	255	260
243	261	249	230	246	263	235	229
218	238	254	249	250	263	229	221
253	227	270	257	261	238	240	239
273	220	226	239	258	259	230	262
255	226						

2.10 The following data represent the number of passengers per flight in a sample of 50 flights from Wichita, Kansas, to Kansas City, Missouri.

23	46	66	67	13	58	19	17	65	17
25	20	47	28	16	38	44	29	48	29
69	34	35	60	37	52	80	59	51	33
48	46	23	38	52	50	17	57	41	77
45	47	49	19	32	64	27	61	70	19

a. Construct a dot plot for these data.

b. Construct a stem-and-leaf plot for these data. What does the stem-and-leaf plot tell you about the number of passengers per flight?

2.3 QUALITATIVE DATA GRAPHS

In contrast to quantitative data graphs that are plotted along a numerical scale, qualitative graphs are plotted using non-numerical categories. In this section, we will examine three types of qualitative data graphs: (1) pie charts, (2) bar charts, and (3) Pareto charts.

Pie Charts

A **pie chart** is *a circular depiction of data where the area of the whole pie represents 100% of the data and slices of the pie represent a percentage breakdown of the sublevels.* Pie charts show the relative magnitudes of the parts to the whole. They are widely used in business, particularly to depict such things as budget categories, market share, and time/resource allocations. However, the use of pie charts is minimized in the sciences and technology because pie charts can lead to less accurate judgments than are possible with other types of graphs.* Generally, it is more difficult for the viewer to interpret the relative size of angles in a pie chart than to judge the length of rectangles in a bar chart. In the feature, Statistics in Business Today, "Where Are Soft Drinks Sold?" graphical depictions of the percentage of sales by place are displayed by both a pie chart and a vertical bar chart.

Construction of the pie chart begins by determining the proportion of the subunit to the whole. Table 2.6 contains annual sales for the top petroleum refining companies in the

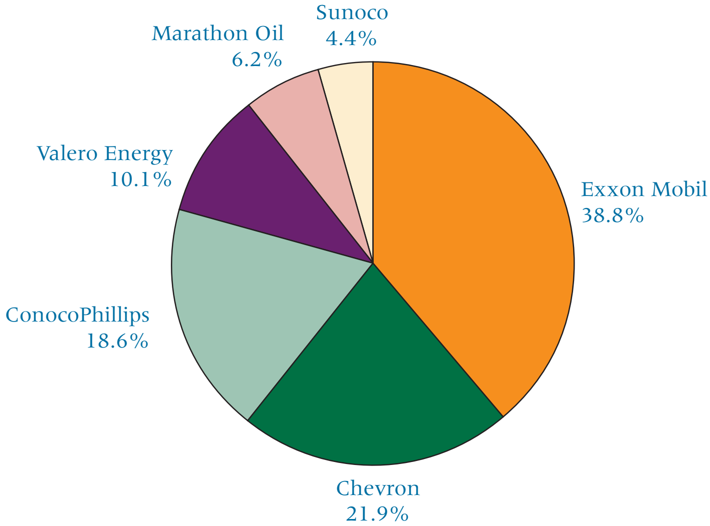
TABLE 2.6
Leading Petroleum Refining Companies

Company	Annual Sales ($ millions)	Proportion	Degrees
Exxon Mobil	372,824	.3879	139.64
Chevron	210,783	.2193	78.95
Conoco Phillips	178,558	.1858	66.89
Valero Energy	96,758	.1007	36.25
Marathon Oil	60,044	.0625	22.50
Sunoco	42,101	.0438	15.77
Totals	961,068	1.0000	360.00

*William S. Cleveland, *The Elements of Graphing Data.* Monterey, CA: Wadsworth Advanced Books and Software, 1985.

FIGURE 2.8

Minitab Pie Chart of Petroleum Refining Sales by Brand



United States in a recent year. To construct a pie chart from these data, first convert the raw sales figures to proportions by dividing each sales figure by the total sales figure. This proportion is analogous to relative frequency computed for frequency distributions. Because a circle contains 360°, each proportion is then multiplied by 360 to obtain the correct number of degrees to represent each item. For example, Exxon Mobil sales of $372,824 million represent a .3879 proportion of the total sales $\left(\frac{372{,}824}{961{,}068} = .3879\right)$. Multiplying this value by 360° results in an angle of 139.64°. The pie chart is then completed by determining each of the other angles and using a compass to lay out the slices. The pie chart in Figure 2.8, constructed by using Minitab, depicts the data from Table 2.6.

Bar Graphs

Another widely used qualitative data graphing technique is the **bar graph** or **bar chart**. A bar graph or chart contains two or more categories along one axis and a series of bars, one for each category, along the other axis. Typically, the length of the bar represents the magnitude of the measure (amount, frequency, money, percentage, etc.) for each category. The bar graph is qualitative because the categories are non-numerical, and it may be either horizontal or vertical. In Excel, horizontal bar graphs are referred to as **bar charts**, and vertical bar graphs are referred to as **column charts**. A bar graph generally is constructed from the same type of data that is used to produce a pie chart. However, an advantage of using a bar graph over a pie chart for a given set of data is that for categories that are close in value, it is considered easier to see the difference in the bars of bar graph than discriminating between pie slices.

As an example, consider the data in Table 2.7 regarding how much the average college student spends on back-to-college spending. Constructing a bar graph from these data, the

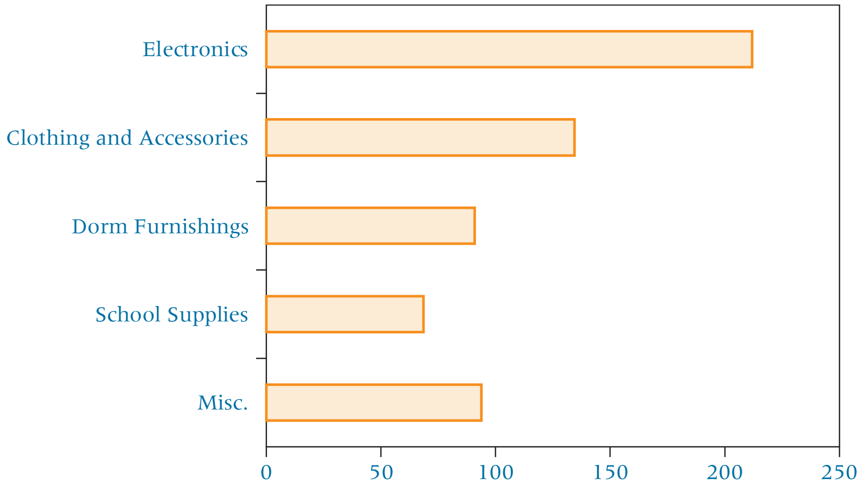
TABLE 2.7

How Much is Spent on Back-to-College Shopping by the Average Student

Category	Amount Spent ($ US)
Electronics	$211.89
Clothing and Accessories	134.40
Dorm Furnishings	90.90
School Supplies	68.47
Misc.	93.72

FIGURE 2.9

Bar Graph of Back-to-College Spending



categories are Electronics, Clothing and Accessories, Dorm Furnishings, School Supplies, and misc. Bars for each of these categories are made using the dollar figures given in the table. The resulting bar graph is shown in Figure 2.9 produced by Excel.

DEMONSTRATION PROBLEM 2.3

According to the National Retail Federation and Center for Retailing Education at the University of Florida, the four main sources of inventory shrinkage are employee theft, shoplifting, administrative error, and vendor fraud. The estimated annual dollar amount in shrinkage ($ millions) associated with each of these sources follows:

Employee theft	$17,918.6
Shoplifting	15,191.9
Administrative error	7,617.6
Vendor fraud	2,553.6
Total	$43,281.7

Construct a pie chart and a bar chart to depict these data.

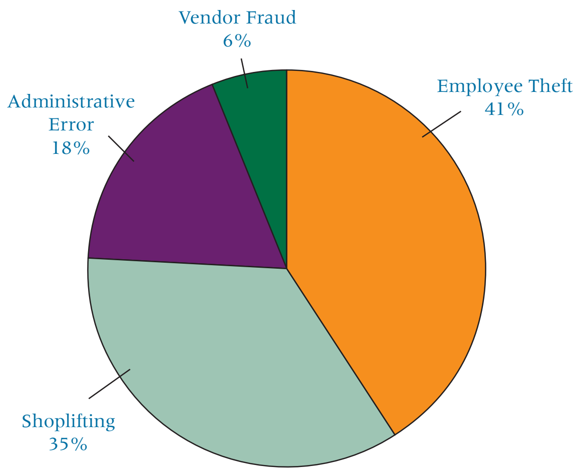
Solution

To produce a pie chart, convert each raw dollar amount to a proportion by dividing each individual amount by the total.

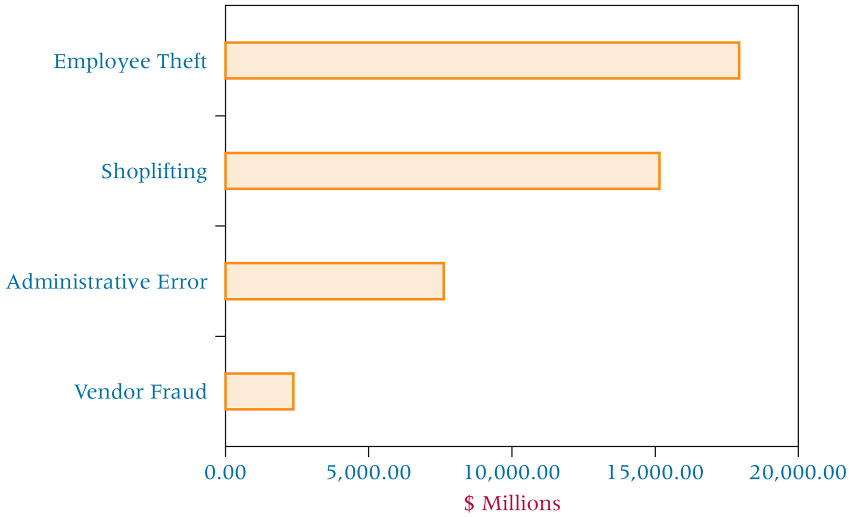
Employee theft	17,918.6/43,281.7 =	.414
Shoplifting	15,191.9/43,281.7 =	.351
Administrative error	7,617.6/43,281.7 =	.176
Vendor fraud	2,553.6/43,281.7 =	.059
Total		1.000

Convert each proportion to degrees by multiplying each proportion by 360°.

Employee theft	.414 · 360° =	149.0°
Shoplifting	.351 · 360° =	126.4°
Administrative error	.176 · 360° =	63.4°
Vendor fraud	.059 · 360° =	21.2°
Total		360.0°



Using the raw data above, we can produce the following bar chart.



Pareto Charts

A third type of qualitative data graph is a Pareto chart, which could be viewed as a particular application of the bar graph. An important concept and movement in business is total quality management (see Chapter 18). One of the important aspects of total quality management is the constant search for causes of problems in products and processes. A graphical technique for displaying problem causes is Pareto analysis. Pareto analysis is a quantitative tallying of the number and types of defects that occur with a product or service. Analysts use this tally to produce *a vertical bar chart that displays the most common types of defects, ranked in order of occurrence from left to right.* The bar chart is called a **Pareto chart**.

Pareto charts were named after an Italian economist, Vilfredo Pareto, who observed more than 100 years ago that most of Italy's wealth was controlled by a few families who were the major drivers behind the Italian economy. Quality expert J. M. Juran applied this notion to the quality field by observing that poor quality can often be addressed by attacking a few major causes that result in most of the problems. A Pareto chart enables quality-management decision makers to separate the most important defects from trivial defects, which helps them to set priorities for needed quality improvement work.

Suppose the number of electric motors being rejected by inspectors for a company has been increasing. Company officials examine the records of several hundred of the motors in which at least one defect was found to determine which defects occurred more frequently. They find that 40% of the defects involved poor wiring, 30% involved a short in the coil, 25%

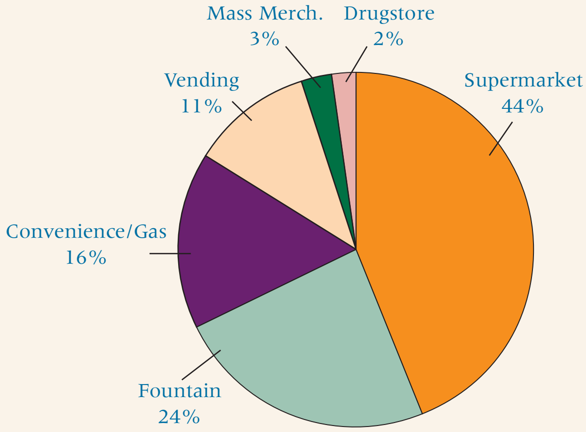
STATISTICS IN BUSINESS TODAY

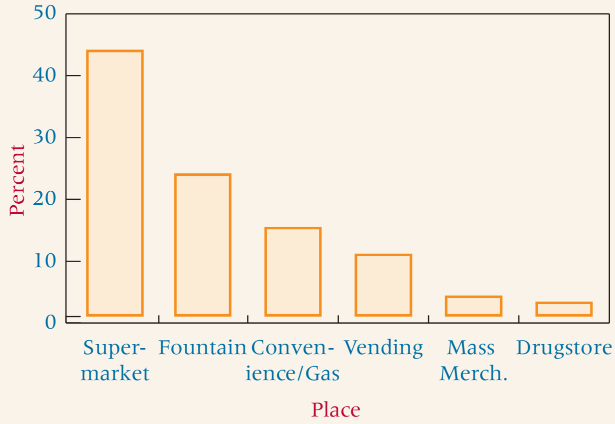
Where Are Soft Drinks Sold?

The soft drink market is an extremely large and growing market in the United States and worldwide. In a recent year, 9.6 billion cases of soft drinks were sold in the United States alone. Where are soft drinks sold? The following data from Sanford C. Bernstein research indicate that the four leading places for soft drink sales are supermarkets, fountains, convenience/gas stores, and vending machines.

Place of Sales	Percentage
Supermarket	44
Fountain	24
Convenience/gas stations	16
Vending	11
Mass merchandisers	3
Drugstores	2

These data can be displayed graphically several ways. Displayed here is an Excel pie chart and a Minitab bar chart of the data. Some statisticians prefer the histogram or the bar chart over the pie chart because they believe it is easier to compare categories that are similar in size with the histogram or the bar chart rather than the pie chart.





involved a defective plug, and 5% involved cessation of bearings. Figure 2.10 is a Pareto chart constructed from this information. It shows that the main three problems with defective motors—poor wiring, a short in the coil, and a defective plug—account for 95% of the problems. From the Pareto chart, decision makers can formulate a logical plan for reducing the number of defects.

Company officials and workers would probably begin to improve quality by examining the segments of the production process that involve the wiring. Next, they would study the construction of the coil, then examine the plugs used and the plug-supplier process.

FIGURE 2.10

Pareto Chart for Electric Motor Problems

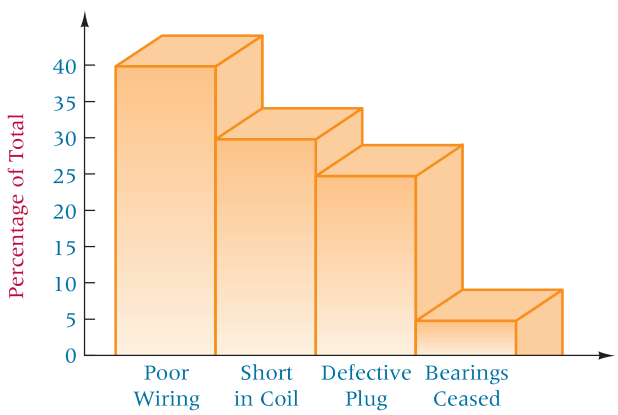


FIGURE 2.11

Minitab Pareto Chart for Electric Motor Problems

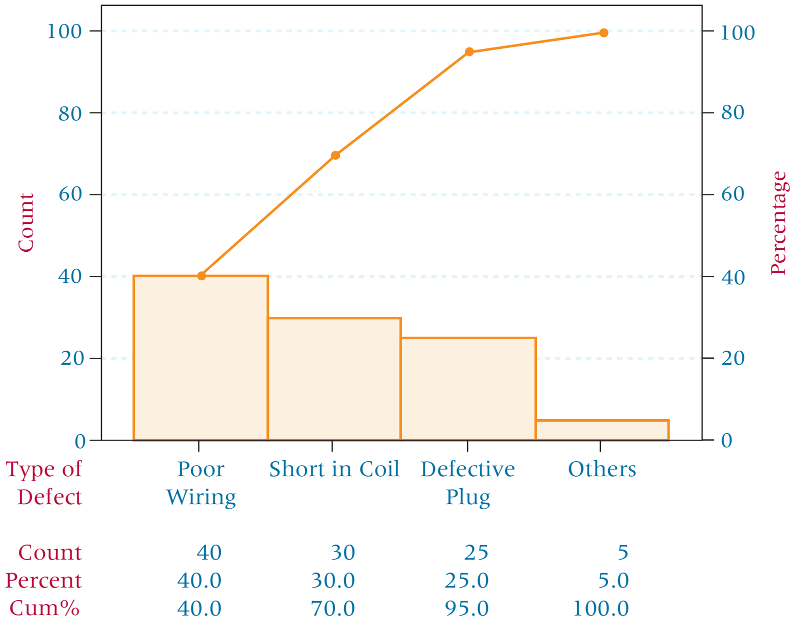


Figure 2.11 is a Minitab rendering of this Pareto chart. In addition to the bar chart analysis, the Minitab Pareto analysis contains a cumulative percentage line graph. Observe the slopes on the line graph. The steepest slopes represent the more frequently occurring problems. As the slopes level off, the problems occur less frequently. The line graph gives the decision maker another tool for determining which problems to solve first.

2.3 PROBLEMS

2.11 Shown here is a list of the top five industrial and farm equipment companies in the United States, along with their annual sales ($ millions). Construct a pie chart and a bar graph to represent these data, and label the slices with the appropriate percentages. Comment on the effectiveness of using a pie chart to display the revenue of these top industrial and farm equipment companies.

Firm	Revenue ($ million)
Caterpillar	30,251
Deere	19,986
Illinois Tool Works	11,731
Eaton	9,817
American Standard	9,509

2.12 According to T-100 Domestic Market, the top seven airlines in the United States by domestic boardings in a recent year were Southwest Airlines with 81.1 million, Delta Airlines with 79.4 million, American Airlines with 72.6 million, United Airlines with 56.3 million, Northwest Airlines with 43.3 million, US Airways with 37.8 million, and Continental Airlines with 31.5 million. Construct a pie chart and a bar graph to depict this information.

2.13 The following list shows the top six pharmaceutical companies in the United States and their sales figures ($ millions) for a recent year. Use this information to construct a pie chart and a bar graph to represent these six companies and their sales.

Pharmaceutical Company	Sales
Pfizer	52,921
Johnson & Johnson	47,348
Merck	22,939
Bristol-Myers Squibb	21,886
Abbott Laboratories	20,473
Wyeth	17,358

2.14 An airline company uses a central telephone bank and a semiautomated telephone process to take reservations. It has been receiving an unusually high number of customer complaints about its reservation system. The company conducted a survey of customers, asking them whether they had encountered any of the following problems in making reservations: busy signal, disconnection, poor connection, too long a wait to talk to someone, could not get through to an agent, connected with the wrong person. Suppose a survey of 744 complaining customers resulted in the following frequency tally.

Number of Complaints	Complaint
184	Too long a wait
10	Transferred to the wrong person
85	Could not get through to an agent
37	Got disconnected
420	Busy signal
8	Poor connection

Construct a Pareto diagram from this information to display the various problems encountered in making reservations.

2.4 GRAPHICAL DEPICTION OF TWO-VARIABLE NUMERICAL DATA: SCATTER PLOTS

Many times in business research it is important to explore the relationship between two numerical variables. A more detailed statistical approach is given in chapter 12, but here we present a graphical mechanism for examining the relationship between two numerical variables—the scatter plot (or scatter diagram). A **scatter plot** is *a two-dimensional graph plot of pairs of points from two numerical variables.*

As an example of two numerical variables, consider the data in Table 2.8. Displayed are the values of new residential and new nonresidential buildings in the United States for various years over a 35-year period. Do these two numerical variables exhibit any relationship? It might seem logical when new construction booms that it would boom in both residential building and in nonresidential building at the same time. However, the Minitab scatter plot of these data displayed in Figure 2.12 shows somewhat mixed results. The apparent tendency is that more new residential building construction occurs when more new nonresidential building construction is also taking place and less new residential building

TABLE 2.8

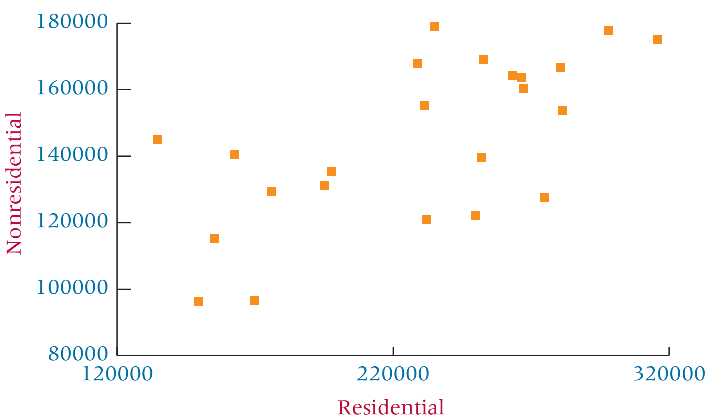
Value of New Construction Over a 35-Year Period

Residential	Nonresidential
169635	96497
155113	115372
149410	96407
175822	129275
162706	140569
134605	145054
195028	131289
231396	155261
234955	178925
266481	163740
267063	160363
263385	164191
252745	169173
228943	167896
197526	135389
232134	120921
249757	122222
274956	127593
251937	139711
281229	153866
280748	166754
297886	177639
315757	175048

Source: U.S. Census Bureau, *Current Construction Reports* (in millions of constant dollars).

FIGURE 2.12

Minitab Scatter Plot of New Residential and New Nonresidential Construction



construction when new nonresidential building construction is also at lower levels. The scatter plot also shows that in some years more new residential building and less new nonresidential building happened at the same time, and vice versa.

2.4 PROBLEMS

2.15 The U.S. National Oceanic and Atmospheric Administration, National Marine Fisheries Service, publishes data on the quantity and value of domestic fishing in the United States. The quantity (in millions of pounds) of fish caught and used for human food and for industrial products (oil, bait, animal food, etc.) over a decade follows. Is a relationship evident between the quantity used for human food and the quantity used for industrial products for a given year? Construct a scatter plot of the data. Examine the plot and discuss the strength of the relationship of the two variables.

Human Food	Industrial Product
3654	2828
3547	2430
3285	3082
3238	3201
3320	3118
3294	2964
3393	2638
3946	2950
4588	2604
6204	2259

2.16 Are the advertising dollars spent by a company related to total sales revenue? The following data represent the advertising dollars and the sales revenues for various companies in a given industry during a recent year. Construct a scatter plot of the data from the two variables and discuss the relationship between the two variables.

Advertising (in $ millions)	Sales (in $ millions)
4.2	155.7
1.6	87.3
6.3	135.6
2.7	99.0
10.4	168.2
7.1	136.9
5.5	101.4
8.3	158.2

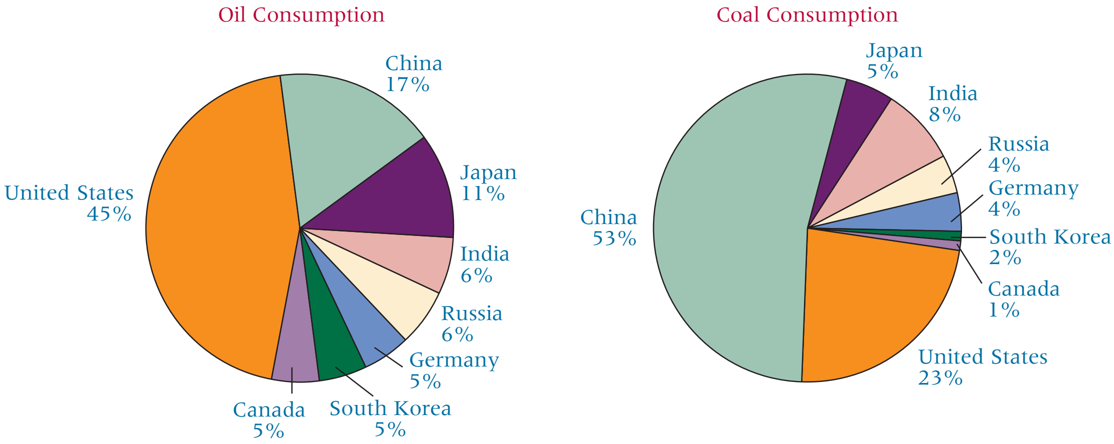
DECISION DILEMMA SOLVED

Energy Consumption Around the World

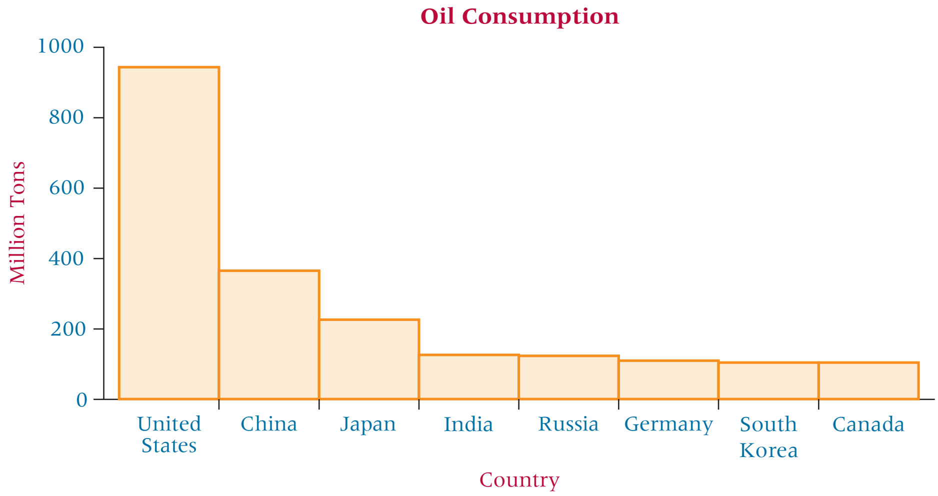
Decision Dilemma Solved

The raw values as shown in the table in the Decision Dilemma are relatively easy to read and interpret. However, these numbers could also be displayed graphically in different ways to create interest and discussion among readers and to allow for more ease of comparisons. For example, shown below are side-by-side Excel pie charts displaying both oil and coal energy consumption figures by country. With such charts, the reader can visually see which countries are dominating consumption of each energy source and then can compare consumption segments across sources.

Pie Charts for World Oil and Coal Consumption (Top Eight Nations)



Sometimes it is difficult for the reader to discern the relative sizes of pie slices that are close in magnitude. For that reason, a bar chart might be a better way to display the data. Shown below is a Minitab-produced histogram of the oil consumption data. It is easy to see that the United States dominates world oil consumption.



ETHICAL CONSIDERATIONS

Ethical considerations for techniques learned in Chapter 2 begin with the data chosen for representation. With the abundance of available data in business, the person constructing the data summary must be selective in choosing the reported variables. The potential is great for the analyst to select variables or even data within variables that are favorable to his or her own situation or that are perceived to be well received by the listener.

Section 2.1 noted that the number of classes and the size of the intervals in frequency distributions are usually selected by the researcher. The researcher should be careful to select values and sizes that will give an honest, accurate reflection of the situation and not a biased over- or under-stated case.

Sections 2.2, 2.3, and 2.4 discussed the construction of charts and graphs. It pointed out that in many instances, it makes sense to use unequal scales on the axes. However, doing so opens the possibility of "cheating with statistics" by stretching or compressing of the axes to underscore the researcher's or analyst's point. It is imperative that frequency distributions and charts and graphs be constructed in a manner that most reflects actual data and not merely the researcher's own agenda.

SUMMARY

The two types of data are grouped and ungrouped. Grouped data are data organized into a frequency distribution. Differentiating between grouped and ungrouped data is important, because statistical operations on the two types are computed differently.

Constructing a frequency distribution involves several steps. The first step is to determine the range of the data, which is the difference between the largest value and the smallest value. Next, the number of classes is determined, which is an arbitrary choice of the researcher. However, too few classes overaggregate the data into meaningless categories, and too many classes do not summarize the data enough to be useful. The third step in constructing the frequency distribution is to determine the width of the class interval. Dividing the range of values by the number of classes yields the approximate width of the class interval.

The class midpoint is the midpoint of a class interval. It is the average of the class endpoints and represents the halfway point of the class interval. Relative frequency is a value computed by dividing an individual frequency by the sum of the frequencies. Relative frequency represents the proportion of total values that is in a given class interval. The cumulative frequency is a running total frequency tally that starts with the first frequency value and adds each ensuing frequency to the total.

Two types of graphical depictions are quantitative data graphs and qualitative data graphs. Quantitative data graphs presented in this chapter are histogram, frequency polygon, ogive, dot plot, and stem-and-leaf plot. Qualitative data graphs presented are pie chart, bar chart, and Pareto chart. In addition, two-dimensional scatter plots are presented. A histogram is a vertical bar chart in which a line segment connects class endpoints at the value of the frequency. Two vertical lines connect this line segment down to the *x*-axis, forming a rectangle. A frequency polygon is constructed by plotting a dot at the midpoint of each class interval for the value of each frequency and then connecting the dots. Ogives are cumulative frequency polygons. Points on an ogive are plotted at the class endpoints. A dot plot is a graph that displays frequency counts for various data points as dots graphed above the data point. Dot plots are especially useful for observing the overall shape of the distribution and determining both gaps in the data and high concentrations of data. Stem-and-leaf plots are another way to organize data. The numbers are divided into two parts, a stem and a leaf. The stems are the leftmost digits of the numbers and the leaves are the rightmost digits. The stems are listed individually, with all leaf values corresponding to each stem displayed beside that stem.

A pie chart is a circular depiction of data. The amount of each category is represented as a slice of the pie proportionate to the total. The researcher is cautioned in using pie charts because it is sometimes difficult to differentiate the relative sizes of the slices.

The bar chart or bar graph uses bars to represent the frequencies of various categories. The bar chart can be displayed horizontally or vertically.

A Pareto chart is a vertical bar chart that is used in total quality management to graphically display the causes of problems. The Pareto chart presents problem causes in descending order to assist the decision maker in prioritizing problem causes. The scatter plot is a two-dimensional plot of pairs of points from two numerical variables. It is used to graphically determine whether any apparent relationship exists between the two variables.

KEY TERMS

Flash Cards

bar graph
class mark
class midpoint
cumulative frequency
dot plot
frequency distribution
frequency polygon
grouped data
histogram
ogive
Pareto chart
pie chart
range
relative frequency
scatter plot
stem-and-leaf plot
ungrouped data

SUPPLEMENTARY PROBLEMS

CALCULATING THE STATISTICS

2.17 For the following data, construct a frequency distribution with six classes.

57	23	35	18	21
26	51	47	29	21
46	43	29	23	39
50	41	19	36	28
31	42	52	29	18
28	46	33	28	20

2.18 For each class interval of the frequency distribution given, determine the class midpoint, the relative frequency, and the cumulative frequency.

Class Interval	Frequency
20–under 25	17
25–under 30	20
30–under 35	16
35–under 40	15
40–under 45	8
45–under 50	6

2.19 Construct a histogram, a frequency polygon, and an ogive for the following frequency distribution.

Class Interval	Frequency
50–under 60	13
60–under 70	27
70–under 80	43
80–under 90	31
90–under 100	9

2.20 Construct a dot plot from the following data.

16	15	17	15	15
15	14	9	16	15
13	10	8	18	20
17	17	17	18	23
7	15	20	10	14

2.21 Construct a stem-and-leaf plot for the following data. Let the leaf contain one digit.

312	324	289	335	298
314	309	294	326	317
290	311	317	301	316
306	286	308	284	324

2.22 Construct a pie chart from the following data.

Label	Value
A	55
B	121
C	83
D	46

2.23 Construct a bar graph from the following data.

Category	Frequency
A	7
B	12
C	14
D	5
E	19

2.24 An examination of rejects shows at least 7 problems. A frequency tally of the problems follows. Construct a Pareto chart for these data.

Problem	Frequency
1	673
2	29
3	108
4	202
5	73
6	564
7	402

2.25 Construct a scatter plot for the following two numerical variables.

x	*y*
12	5
17	3
9	10
6	15
10	8
14	9
8	8

TESTING YOUR UNDERSTANDING

2.26 The Whitcomb Company manufactures a metal ring for industrial engines that usually weighs about 50 ounces. A random sample of 50 of these metal rings produced the following weights (in ounces).

51	53	56	50	44	47
53	53	42	57	46	55
41	44	52	56	50	57
44	46	41	52	69	53
57	51	54	63	42	47
47	52	53	46	36	58
51	38	49	50	62	39
44	55	43	52	43	42
57	49				

Construct a frequency distribution for these data using eight classes. What can you observe about the data from the frequency distribution?

2.27 A northwestern distribution company surveyed 53 of its midlevel managers. The survey obtained the ages of these managers, which later were organized into the frequency distribution shown. Determine the class midpoint, relative frequency, and cumulative frequency for these data.

Class Interval	Frequency
20–under 25	8
25–under 30	6
30–under 35	5
35–under 40	12
40–under 45	15
45–under 50	7

2.28 Use the data from Problem 2.27.
a. Construct a histogram and a frequency polygon.
b. Construct an ogive.

2.29 The following data are shaped roughly like a normal distribution (discussed in Chapter 6).

61.4	27.3	26.4	37.4	30.4	47.5
63.9	46.8	67.9	19.1	81.6	47.9
73.4	54.6	65.1	53.3	71.6	58.6
57.3	87.8	71.1	74.1	48.9	60.2
54.8	60.5	32.5	61.7	55.1	48.2
56.8	60.1	52.9	60.5	55.6	38.1
76.4	46.8	19.9	27.3	77.4	58.1
32.1	54.9	32.7	40.1	52.7	32.5
35.3	39.1				

Construct a frequency distribution starting with 10 as the lowest class beginning point and use a class width of 10. Construct a histogram and a frequency polygon for this frequency distribution and observe the shape of a normal distribution. On the basis of your results from these graphs, what does a normal distribution look like?

2.30 In a medium-sized southern city, 86 houses are for sale, each having about 2000 square feet of floor space. The asking prices vary. The frequency distribution shown contains the price categories for the 86 houses. Construct a histogram, a frequency polygon, and an ogive from these data.

Asking Price	Frequency
$ 80,000–under $100,000	21
100,000–under 120,000	27
120,000–under 140,000	18
140,000–under 160,000	11
160,000–under 180,000	6
180,000–under 200,000	3

2.31 Good, relatively inexpensive prenatal care often can prevent a lifetime of expense owing to complications resulting from a baby's low birth weight. A survey of a random sample of 57 new mothers asked them to estimate how much they spent on prenatal care. The researcher tallied the results and presented them in the frequency distribution shown. Use these data to construct a histogram, a frequency polygon, and an ogive.

Amount Spent on Prenatal Care	Frequency of New Mothers
$ 0–under $100	3
100–under 200	6
200–under 300	12
300–under 400	19
400–under 500	11
500–under 600	6

2.32 A consumer group surveyed food prices at 87 stores on the East Coast. Among the food prices being measured was that of sugar. From the data collected, the group constructed the frequency distribution of the prices of 5 pounds of Domino's sugar in the stores surveyed. Compute a histogram, a frequency polygon, and an ogive for the following data.

Price	Frequency
$1.75–under $1.90	9
1.90–under 2.05	14
2.05–under 2.20	17
2.20–under 2.35	16
2.35–under 2.50	18
2.50–under 2.65	8
2.65–under 2.80	5

2.33 The top music genres according to SoundScan for a recent year are R&B, Alternative (Rock), Rap, and Country. These and other music genres along with the number of albums sold in each (in millions) are shown.

Genre	Albums Sold
R&B	146.4
Alternative	102.6
Rap	73.7
Country	64.5
Soundtrack	56.4
Metal	26.6
Classical	14.8
Latin	14.5

Construct a pie chart for these data displaying the percentage of the whole that each of these genres represents. Construct a bar chart for these data.

2.34 The following figures for U.S. imports of agricultural products and manufactured goods were taken from selected years over a 30-year period (in $ billions). The source of the data is the U.S. International Trade Administration. Construct a scatter plot for these data and determine whether any relationship is apparent between the U.S. imports of agricultural products and the U.S. imports of manufactured goods during this time period.

Agricultural Products	Manufactured Goods
5.8	27.3
9.5	54.0
17.4	133.0
19.5	257.5
22.3	388.8
29.3	629.7

2.35 Shown here is a list of the industries with the largest total release of toxic chemicals in a recent year according to the U.S. Environmental Protection Agency. Construct a pie chart and a bar chart to depict this information.

Industry	Total Release (pounds)
Chemicals	737,100,000
Primary metals	566,400,000
Paper	229,900,000
Plastics and rubber	109,700,000
Transportation equipment	102,500,000
Food	89,300,000
Fabricated metals	85,900,000
Petroleum	63,300,000
Electrical equipment	29,100,000

2.36 A manufacturing company produces plastic bottles for the dairy industry. Some of the bottles are rejected because of poor quality. Causes of poor-quality bottles include faulty plastic, incorrect labeling, discoloration, incorrect thickness, broken handle, and others. The following data for 500 plastic bottles that were rejected include the problems and the frequency of the problems. Use these data to construct a Pareto chart. Discuss the implications of the chart.

Problem	Number
Discoloration	32
Thickness	117
Broken handle	86
Fault in plastic	221
Labeling	44

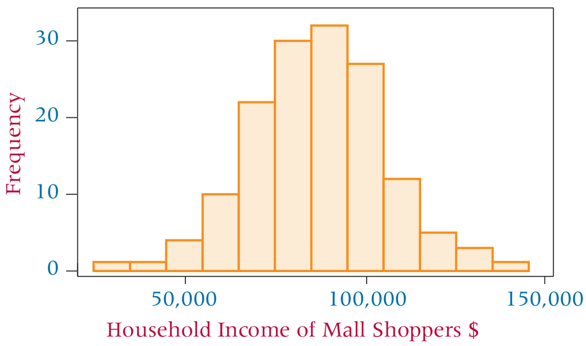
2.37 A research organization selected 50 U.S. towns with Census 2000 populations between 4,000 and 6,000 as a sample to represent small towns for survey purposes. The populations of these towns follow.

4420	5221	4299	5831	5750
5049	5556	4361	5737	4654
4653	5338	4512	4388	5923
4730	4963	5090	4822	4304
4758	5366	5431	5291	5254
4866	5858	4346	4734	5919
4216	4328	4459	5832	5873
5257	5048	4232	4878	5166
5366	4212	5669	4224	4440
4299	5263	4339	4834	5478

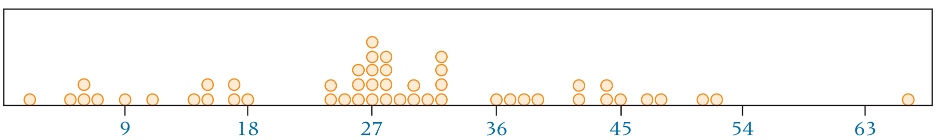
Construct a stem-and-leaf plot for the data, letting each leaf contain two digits.

INTERPRETING THE OUTPUT

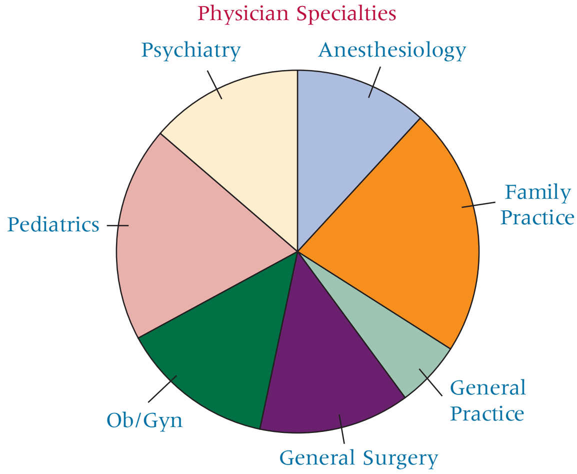
2.38 Suppose 150 shoppers at an upscale mall are interviewed and one of the questions asked is the household income. Study the Minitab histogram of the following data and discuss what can be learned about the shoppers.



2.39 Study the following dot plot and comment on the general shape of the distribution. Discuss any gaps or heavy concentrations in the data.



2.40 Shown here is an Excel-produced pie chart representing physician specialties. What does the chart tell you about the various specialties?



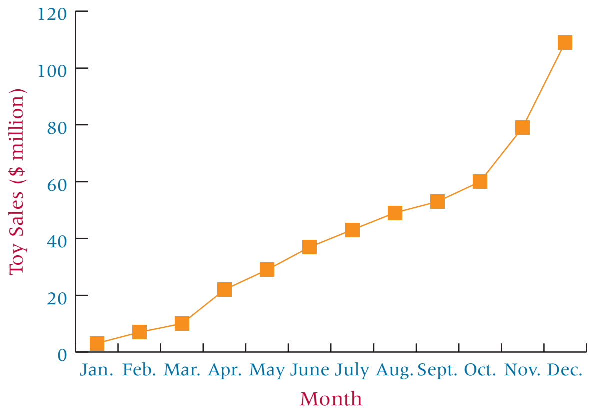
2.41 Suppose 100 CPA firms are surveyed to determine how many audits they perform over a certain time. The data are summarized using the Minitab stem-and-leaf plot shown in the next column. What can you learn about the number of audits being performed by these firms from this plot?

Stem-and-Leaf Display: Audits

Stem-and-leaf of Audits N = 100
Leaf Unit = 1.0

9	1	222333333
16	1	4445555
26	1	6666667777
35	1	888899999
39	2	0001
44	2	22333
49	2	55555
(9)	2	677777777
42	2	8888899
35	3	000111
29	3	223333
23	3	44455555
15	3	67777
10	3	889
7	4	0011
3	4	222

2.42 The following Excel ogive shows toy sales by a company over a 12-month period. What conclusions can you reach about toy sales at this company?



ANALYZING THE DATABASES

see www.wiley.com/college/black

Database

1. Using the manufacturer database, construct a frequency distribution for the variable Number of Production Workers. What does the frequency distribution reveal about the number of production workers?

2. Using the Consumer Food database, construct a histogram for the variable Annual Food Spending. How is the histogram shaped? Is it high in the middle or high near one or both ends of the data? Is it relatively constant in size across the class (uniform), or does it appear to have no shape? Does it appear to be nearly "normally" distributed?

3. Construct an ogive for the variable Type in the financial database. The 100 companies in this database are each categorized into one of seven types of companies. These types are listed at the end of Chapter 1. Construct a pie chart of these types and discuss the output. For example, which type is most prevalent in the database and which is the least?

4. Using the international unemployment database, construct a stem-and-leaf plot for Italy. What does the plot show about unemployment for Italy over the past 40 years? What does the plot fail to show?

CASE

SOAP COMPANIES DO BATTLE

Procter & Gamble has been the leading soap manufacturer in the United States since 1879, when it introduced Ivory soap. However, late in 1991, its major rival, Lever Bros. (Unilever), overtook it by grabbing 31.5% of the $1.6 billion personal soap market, of which Procter & Gamble had a 30.5% share. Lever Bros. had trailed Procter & Gamble since it entered the soap market with Lifebuoy in 1895. In 1990, Lever Bros. introduced a new soap, Lever 2000, into its product mix as a soap for the entire family. A niche for such a soap had been created because of the segmentation of the soap market into specialty soaps for children, women, and men. Lever Bros. felt that it could sell a soap for everyone in the family. Consumer response was strong; Lever 2000 rolled up $113 million in sales in 1991, putting Lever Bros. ahead of Procter & Gamble for the first time in the personal-soap revenue contest. Procter & Gamble still sells more soap, but Lever's brands cost more, thereby resulting in greater overall sales.

Needless to say, Procter & Gamble was quick to search for a response to the success of Lever 2000. Procter & Gamble looked at several possible strategies, including repositioning Safeguard, which has been seen as a male soap. Ultimately, Procter & Gamble responded to the challenge by introducing its Oil of Olay Moisturizing Bath Bar. In its first year of national distribution, this product was backed by a $24 million media effort. The new bath bar was quite successful and helped Procter & Gamble regain market share.

These two major companies continue to battle it out for domination in the personal soap market, along with the Dial Corporation and Colgate-Palmolive.

Shown below are sales figures in a recent year for personal soaps in the United States. Each of these soaps is produced by one of four soap manufacturers: Unilever, Procter & Gamble, Dial, and Colgate-Palmolive.

Soap	Manufacturer	Sales ($ millions)
Dove	Unilever	271
Dial	Dial	193
Lever 2000	Unilever	138
Irish Spring	Colgate-Palmolive	121
Zest	Procter & Gamble	115
Ivory	Procter & Gamble	94
Caress	Unilever	93
Olay	Procter & Gamble	69
Safeguard	Procter & Gamble	48
Coast	Dial	44

In 1983, the market shares for soap were Procter & Gamble with 37.1%, Lever Bros. (Unilever) with 24%, Dial with 15%, Colgate-Palmolive with 6.5%, and all others with 17.4%. By 1991, the market shares for soap were Lever Bros. (Unilever) with 31.5%, Procter & Gamble with 30.5%, Dial with 19%, Colgate-Palmolive with 8%, and all others with 11%.

Discussion

1. Suppose you are making a report for Procter & Gamble displaying their share of the market along with the share of other companies for the years 1983, 1991, and the latest figures. Using either Excel or Minitab, produce graphs for the market shares of personal soap for each of these years. For the latest figures data, assume that the "all others" total is about $119 million. What do you observe about the market shares of the various companies by studying the graphs? In particular, how is Procter & Gamble doing relative to previous years?
2. Suppose Procter & Gamble sells about 20 million bars of soap per week, but the demand is not constant and production management would like to get a better handle on how sales are distributed over the year. Let the following sales figures given in units of million bars represent the sales of bars per week over one year. Construct a histogram to represent these data. What do you see in the graph that might be helpful to the production (and sales) people?

17.1	19.6	15.4	17.4	15.0	18.5	20.6	18.4
20.0	20.9	19.3	18.2	14.7	17.1	12.2	19.9
18.7	20.4	20.3	15.5	16.8	19.1	20.4	15.4
20.3	17.5	17.0	18.3	13.6	39.8	20.7	21.3
22.5	21.4	23.4	23.1	22.8	21.4	24.0	25.2
26.3	23.9	30.6	25.2	26.2	26.9	32.8	26.3
26.6	24.3	26.2	23.8				

Construct a stem-and-leaf plot using the whole numbers as the stems. What advantages does the stem-and-leaf plot of these sales figures offer over the histogram? What are some disadvantages? Which would you use in discussions with production people, and why?

3. A random sample of finished soap bars in their packaging is tested for quality. All defective bars are examined for problem causes. Among the problems found were improper packaging, poor labeling, bad seal, shape of bar wrong, bar surface marred, wrong color in bar, wrong bar fragrance, wrong soap consistency, and others. Some of the leading problem causes and the number of each are given here. Use a Pareto chart to analyze these problem causes. Based on your findings, what would you recommend to the company?

Problem Cause	Frequency
Bar surface	89
Color	17
Fragrance	2
Label	32
Shape	8
Seal	47
Labeling	5
Soap consistency	3

Source: Adapted from Valerie Reitman, "Buoyant Sales of Lever 2000 Soap Bring Sinking Sensation to Procter & Gamble," *The Wall Street Journal*, March 19, 1992, p. B1. Reprinted by permission of *The Wall Street Journal* © 1992, Dow Jones & Company, Inc. All rights reserved worldwide; Pam Weisz, "$40 M Extends Lever 2000 Family," *Brandweek*, vol. 36, no. 32 (August 21, 1995), p. 6; Laurie Freeman, "P&G Pushes Back Against Unilever in Soap," *Advertising Age*, vol. 65, no. 41 (September 28, 1994), p. 21; Jeanne Whalen and Pat Sloan, "Intros Help Boost Soap Coupons," *Advertising Age*, vol. 65, no. 19 (May 2, 1994), p. 30; and "P&G Places Coast Soap up for Sale," *The Post*, World Wide Web Edition of *The Cincinnati Post*, February 2, 1999, http://www.cincypost.com.business /pg022599.html.

USING THE COMPUTER

EXCEL

- Excel offers the capability of producing many of the charts and graphs presented in this chapter. Most of these can be accessed by clicking on the **Insert** tab found along the top of an Excel worksheet (second tab from the left next to **Home**). In addition, Excel can generate frequency distributions and histograms using the **Data Analysis** feature.
- Many of the statistical techniques presented in this text can be performed in Excel using a tool called **Data Analysis**. To access this feature, select the **Data** tab along the top of an Excel worksheet. The **Data** tab is the fifth tab over from the left. If the **Data Analysis** feature has been uploaded into your Excel package, it will be found in the **Analysis** section at the top right of the **Data** tab page on the far right. If **Data Analysis** does not appear in the **Analysis** section, it must be added in. To add in **Data Analysis**: 1.) Click on the Microsoft Office logo button located in the very topmost left of the Excel 2007 page (looks like an office icon and is called **Office Button**). 2.) Now click on **Excel Options** located at the bottom of the pulldown menu. 3.) From the menu of the left panel of the **Excel Options** dialog box, click on **Add-Ins**. From the resulting menu shown on the right side of the dialog box, highlight **Analysis ToolPak**. Click on **Go…** at the bottom of the page. An **Add-Ins** dialog box will appear with a menu. Check **Analysis ToolPak** and click on **OK**. Your **Data Analysis** feature is now uploaded onto your computer, and you need not add it in again. Now you can bring up the **Analysis ToolPak** feature at any time by going to the **Data** tab at the top of the Excel worksheet and clicking on **Data Analysis**.
- In Excel, frequency distributions are referred to as histograms, and the classes of a frequency distribution are referred to as bins. If you do not specify bins (classes), Excel will automatically determine the number of bins and assign class endpoints based on a formula. If you want to specify bins, load the class endpoints that you want to use

into a column. To construct a frequency distribution, select the **Data** tab in the Excel worksheet and then select the **Data Analysis** feature (upper right). If this feature does not appear, you may need to add it (see above). Clicking on **Data Analysis**, the dialog box features a pulldown menu of many of the statistical analysis tools presented and used in this text. From this list, select **Histogram**. In the **Histogram** dialog box, place the location of the raw data values in the space beside **Input Range**. Place the location of the class endpoints (optional) in the space beside **Bin Range**. Leave this blank if you want Excel to determine the bins (classes). If you have labels, check **Labels**. If you want a histogram graph, check **Chart Output**. If you want an ogive, select **Cumulative Percentage** along with **Chart Output**. If you opt for this, Excel will yield a histogram graph with an ogive overlaid on it.

- Excel has excellent capability of constructing many different types of charts, including column charts, line charts, pie charts, bar charts, area charts, and XY (scatter) charts. To begin the process of producing these charts, select the **Insert** tab from the top of the Excel 2007 worksheet. In the **Charts** section, which is the middle section shown at the top of the **Insert** worksheet, there are icons for column, line, pie, bar, area, scatter, and other charts. Click on the icon representing the desired chart to begin construction. Each of these types of charts allow for several versions of the chart shown in the dropdown menu. For example, the pie chart menu contains four types of two-dimensional pie charts and two types of three-dimensional pie charts. To select a particular version of a type of chart, click on the type of chart and then the version of that chart that is desired.
- To construct a pie chart, enter the categories in one column and the data values of each category in another column in the Excel worksheet. Categories and data values could also be entered in rows instead of columns. Click and drag over the data for which the pie chart is to be constructed. From the **Insert** tab, select **Pie** from the **Charts** section and the select the type of pie chart to be constructed. The result is a pie chart from the data. Once the chart has been constructed, a set of three new tabs appear at the top of the worksheet under the general area of **Chart Tools** (see top upper right corner of worksheet). The three new tabs are **Design**, **Layout**, and **Format**. There are many options available for changing the design of the pie chart that can be accessed by clicking on the up and down arrow on the right end of the **Design** tab in the section called **Chart Styles**. On the far right end of the **Design** menu bar is a feature called **Move Chart Location**, which can be used to move the chart to another location or to a new sheet. On the far left end of the **Design** menu bar, there is a **Change Chart Type** feature that allows for changing the type of chart that has been constructed. The second group of features from the left at the top of the **Design** menu bar makes it possible to switch to another set of data (**Select Data**) or switch rows and columns (**Switch Row/Column**). There is a useful feature in the middle of the Design menu bar called **Quick Layout** that offers several different layouts for the given chart type. For example, for pie charts, there are seven different possible layouts using titles, labels, and legends in different ways. Right-clicking while on the pie chart brings up a menu that includes **Format Data Labels** and **Format Data Series**. Clicking on **Format Data Labels** brings up another menu (shown on the left) that allows you to modify or edit various features of your graph, including **Label Options**, **Number**, **Fill**, **Border Color**, **Border Styles**, **Shadow**, **3-D Format**, and **Alignment**. Under **Label Options**, there are several different types of labels for pie charts and there are other various chart options available, such as **Series Name**, **Category Name**, **Value**, **Percentage**, and **Show Leader Lines**. In addition, it offers various options for the label location, such as **Center**, **Inside End**, **Outside End**, and **Best Fit**. It also offers the opportunity to include the legend key in the label. The **Number** option under **Format Data Labels…** allows for the usual Excel options in using numbers. The **Fill** option allows you to determine what type of fill you want to have for the chart. Options include **No fill**, **Solid fill**, **Gradient fill**, **Picture or texture fill**, and **Automatic**. Other options under **Format Data Labels…** allow you to manipulate the border colors and styles, shadow, 3-D format, and text alignment or layout. The **Layout** tab at the top of the worksheet page has a **Labels** panel located at the top of the worksheet page just to the left of the center. In this section, you can further specify the location of the chart title by selecting **Chart Title**, the location of the legend by selecting **Legend**, or the location of the labels by selecting **Data Labels**. The **Format** tab at the top of the worksheet page contains a **Shape Styles** panel just to the left of center at the top of the worksheet. This panel contains options for visual styles of the graph (for more options, use the up and down arrow) and options for **Shape Fill**, **Shape Outline**, and **Shape Effects**. Other formatting options are available through the use of the **Format Selection** option on the far upper left of the **Current Selection** panel on the **Format** tab page.
- Frequency polygons can be constructed in Excel 2007 by using the **Histogram** feature. Follow the directions shown above to construct a histogram. Once the histogram is constructed, right-click on one of the “bars” of the histogram. From the dropdown menu, select **Change Series Chart Type**. Next select a line chart type. The result will be a frequency polygon.
- An ogive can be constructed at least two ways. One way is to cumulate the data manually. Enter the cumulated data in one column and the class endpoints in another column. Click and drag over both columns. Go to the **Insert** tab at the top of the Excel worksheet. Select **Scatter** as the type of chart. Under the **Scatter** options, select the option with the solid lines. The result is an ogive. A second way is to construct a frequency distribution first using the **Histogram** feature in the **Data Analysis** tool. In the **Histogram** dialog box, enter the location of the data and enter the location of

the class endpoints as bin numbers. Check **Cumulative Percentage** and **Chart Output** in the **Histogram** dialog box. Once the chart is constructed, right-click on one of the bars and select the **Delete** option. The result will be an ogive chart with just the ogive line graph (and bars eliminated).

- Bar charts and column charts are constructed in a manner similar to that of a pie chart. Begin by entering the categories in one column and the data values of each category in another column in the Excel worksheet. Categories and data values could also be entered in rows instead of columns. Click and drag over the data and categories for which the chart is to be constructed. Go to the **Insert** tab at the top of the worksheet. Select **Column** or **Bar** from the **Charts** section and the select the version of the chart to be constructed. The result is a chart from the data. Once the bar chart or column chart has been constructed, there are many options available to you. By right-clicking on the bars or columns, a menu appears that allows you, among other things, to label the columns or bars. This command is **Add Data Labels**. Once data labels are added, clicking on the bars or columns will allow you to modify the labels and the characteristics of the bars or columns by selecting **Format Data Labels…** or **Format Data Series….** Usage of these commands is the same as when constructing or modifying pie charts (see above). Various options are also available under **Chart Tools** (see pie charts above).
- Pareto charts, as presented in the text, have categories and numbers of defects. As such, Pareto charts can be constructed as **Column** charts in Excel using the same commands (see above). However, the user will first need to order the categories and their associated frequencies in descending order. In addition, in constructing a histogram in Excel (see above), there is an option in the **Histogram** dialog box called **Pareto (sorted histogram)** in which Excel takes histogram data and presents the data with categories organized from highest frequency to lowest.
- Scatter diagrams can be constructed in Excel. Begin by entering the data for the two variables to be graphed in two separate rows or columns. You may either use a label for each variable or not. Click and drag over the data (and labels). Go to the **Insert** tab. From the **Charts** panel (upper middle), select **Scatter**. From the ensuing pulldown menu of scatter plot options, select one of the versions from the five presented. The result is the scatter chart. By right-clicking on the chart, various other chart options are available including, **Format Plot Area….** The resulting menu associated with this command offers the usual chart options regarding fill, border color, border styles, shadow, and 3-D format (see pie charts above). In addition, if you want to fit a line or curve to the data, right-click on one of the chart points. A menu pops up containing, among other options, **Add Trendline….** From the **Trendline Options**, select the type of line or curve that you want to fit to the data. The result is a line or curve shown on the scatter plot attempting to fit to the points. Various other options are available regarding the line color, style, and shadow.

MINITAB

- Minitab has the capability of constructing histograms, dot plots, stem-and-leaf charts, pie charts, bar charts, and Pareto charts. With the exception of Pareto charts, which are accessed through **Stat**, all of these other charts and graphs are accessed by selecting **Graph** on the menu bar.
- To construct a histogram, select **Graph** on the Minitab menu bar, and then select **Histogram.** The first dialog box offers four histogram options: **Simple, With Fit, With Outline and Groups**, and **With Fit and Groups**. Select the **Simple** option, which is also the default option. In the dialog box that follows beside **Graph variables**, insert the column location (or columns) containing the data for which you want to create a histogram. There are several options from which to modify the histogram. Select **Scale** to adjust or modify the axes, ticks, and gridlines. Select **Labels** to title the graph and label the axes. Select **Data view** for optional ways to present the data, including bars, symbols, project lines, and areas in addition to presenting other options such as smoothing the data. Select **Multiple graphs** to create multiple separate graphs or to combine several graphs on one. Select **Data options** for several options in grouping data.
- To construct a dot plot, select **Graph** on the Minitab menu bar, and then select **Dotplot**. The first dialog box offers seven different ways to configure the plot. Select **Simple** to produce a dot plot like those shown in this chapter. In the dialog box that follows, insert the column location(s) containing the data for which you want to create the dot plot in **Graph variables**. There are several options available. Select **Scale** to adjust or modify the axes and ticks. Select **Labels** to title the graph and add footnotes. Select **Multiple graphs** to create multiple separate graphs or to combine several graphs on one. Select **Data options** for options in grouping data, frequencies, and subsets.
- To construct a stem-and-leaf chart, select **Stem-and-Leaf** from the **Graph** pulldown menu. In the **Stem-and-Leaf** dialog box, place the name of the column(s) containing the data in the **Graph variables** space. Click **OK** and the stem-and-leaf plot is generated. If you have a grouping variable in another column and want to use it, enter the location or name of the column into the **By variable** space. You can trim outliers from the plot by checking **Trim outliers.**
- To construct a pie chart, select **Graph** on the Minitab menu bar, and then select **PieChart** on the **Graph** pulldown menu. In the **Pie Chart** dialog box, there are two options to consider: **Chart counts of unique values** and **Chart values from a table.** If you have raw data and you want Minitab to group them for you into a pie chart, select **Chart counts of unique values.** You can also use this command even if you have text data. On the other hand, if your data are in one column and your categories are in another column, select **Chart values from a table.** The dialog box will ask for the name of the **Categorical variable** and the name of the **Summary variables.** Several options are available to

modify the pie chart, including **Labels, Multiple graphs**, and **Data options.** Several **Pie Options** are available, including how the pie slices are ordered, the starting angle, and the option of combining small categories.

- To construct a bar chart, select **Graph** on the Minitab menu bar, then select **Bar Chart**. In the **Bar Chart** dialog box, there are three options available. To construct a bar chart like those presented in the chapter, select **Simple**. In the dialog box that follows, enter the column(s) containing the data in **Categorical variables**. Several options are available to modify the bar chart, including **Chart Options, Labels**, **Scale**, **Data View**, **Multiple Graphs**, and **Data Options**.
- To construct a Pareto chart, select **Stat** from the menu bar, and then from the pulldown menu that appears, select **Quality Tools.** From the **Quality Tools** pulldown menu, select **Pareto Chart.** From the **Pareto Chart** dialog box, select **Chart defects table** if you have a summary of the defects with the reasons (**Labels in**) in one column and the frequency of occurrence (**Frequencies in**) in another column. Enter the location of the reasons in **Labels in** and the location of the frequencies in **Frequencies in.** If you have unsummarized data, you can select **Chart defects data in.** In the space provided, give the location of the column with all the defects that occurred. It is possible to have the defects either by name or with some code. If you want to have the labels in one column and the defects in another, then select **By variable in** and place the location of the labels there.
- To construct a scatter plot, select **Graph**, then select **Scatterplot**. In the **Scatterplot** dialog box, select the type of scatter plot you want from **Simple, With Groups, With Regression, With Regression and Groups, With Connect Line**, and **With Connect and Groups**. In the second dialog box, enter the *x* and *y* variable names/locations. There are several options from which to modify the scatter plot. Select **Scale** to adjust or modify the axes, ticks, and gridlines. Select **Labels** to title the graph and label the axes. Select **Data view** for optional ways to present the data, including bars, symbols, project lines, and areas in addition to presenting other options such as smoothing the data. Select **Multiple graphs** to create multiple separate graphs or to combine several graphs on one. Select **Data options** for several options in grouping data.