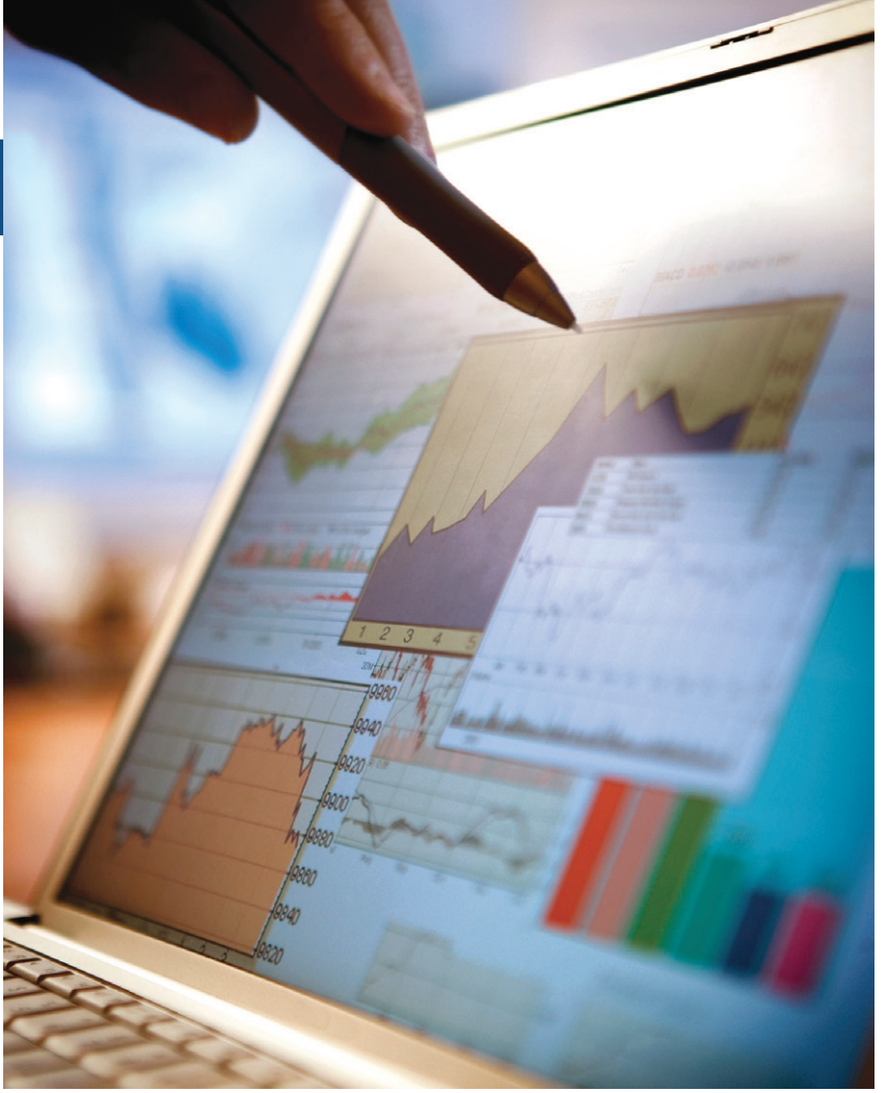
CHAPTER 2

Descriptive Statistics: Tabular and Graphical Methods



Learning Objectives

When you have mastered the material in this chapter, you will be able to:

- LO2-1 Summarize qualitative data by using frequency distributions, bar charts, and pie charts.
- LO2-2 Construct and interpret Pareto charts (Optional).
- LO2-3 Summarize quantitative data by using frequency distributions, histograms, frequency polygons, and ogives.
- LO2-4 Construct and interpret dot plots.
- LO2-5 Construct and interpret stem-and-leaf displays.
- LO2-6 Examine the relationships between variables by using contingency tables (Optional).
- LO2-7 Examine the relationships between variables by using scatter plots (Optional).
- LO2-8 Recognize misleading graphs and charts (Optional).

Chapter Outline

2.1 Graphically Summarizing Qualitative Data
2.2 Graphically Summarizing Quantitative Data
2.3 Dot Plots
2.4 Stem-and-Leaf Displays
2.5 Contingency Tables (Optional)
2.6 Scatter Plots (Optional)
2.7 Misleading Graphs and Charts (Optional)

In Chapter 1 we saw that although we can sometimes take a census of an entire population, we often must randomly select a sample from a population. When we have taken a census or a sample, we typically wish to describe the observed data set. In particular, we describe a sample in order to make inferences about the sampled population.

In this chapter we begin to study **descriptive statistics,** which is the science of describing the important characteristics of a data set. The techniques of descriptive statistics include **tabular and graphical methods,** which are discussed in this chapter, and **numerical methods,** which are discussed in Chapter 3. We will see that, in practice, the methods of this chapter and the methods of Chapter 3 are used together to describe data. We will also see that the methods used to describe quantitative data differ somewhat from the methods used to describe qualitative data. Finally, we will see that there are methods—both graphical and numerical—for studying the relationships between variables.

We will illustrate the methods of this chapter by describing the cell phone usages, bottle design ratings, and car mileages introduced in the cases of Chapter 1. In addition, we introduce two new cases:

The e-billing Case: A management consulting firm assesses how effectively a new electronic billing system reduces bill payment times.

The Brokerage Firm Case: A financial broker examines whether customer satisfaction depends upon the type of investment product purchased.

2.1 Graphically Summarizing Qualitative Data

LO2-1 Summarize qualitative data by using frequency distributions, bar charts, and pie charts.

Frequency distributions When data are qualitative, we use names to identify the different categories (or classes). Often we summarize qualitative data by using a frequency distribution.

A **frequency distribution** is a table that summarizes the number (or **frequency**) of items in each of several nonoverlapping classes.

EXAMPLE 2.1 Describing Pizza Preferences

A business entrepreneur plans to open a pizza restaurant in a college town. There are currently six pizza restaurants in town: four chain restaurants—Domino's Pizza, Little Caesars Pizza, Papa John's Pizza, and Pizza Hut—and two local establishments—Bruno's Pizza and Will's Uptown Pizza. Before developing a basic pizza recipe (crust ingredients, sauce ingredients, and so forth), the entrepreneur wishes to study the pizza preferences of the college students in town. In order to do this, the entrepreneur selects a random sample of 50 students enrolled in the local college and asks each sampled student to name his or her favorite among the six pizza places in town. The survey results are given in Table 2.1.

Part 1: Studying pizza preferences by using a frequency distribution Unfortunately, the raw data in Table 2.1 do not reveal much useful information about the pattern of pizza preferences. In order to summarize the data in a more useful way, we can construct a frequency distribution. To do this we simply count the number of times each of the six pizza restaurants appears in Table 2.1. We find that Bruno's appears 8 times, Domino's appears 2 times, Little Caesars appears 9 times, Papa John's appears 19 times, Pizza Hut appears 4 times, and Will's Uptown Pizza appears 8 times. The frequency distribution for the pizza preferences is given in Table 2.2—a list of each of the six restaurants along with their corresponding counts (or **frequencies**). The frequency distribution shows us how the preferences are distributed among the six restaurants. The purpose of the frequency distribution is to make the data easier to understand. Certainly, looking at the frequency distribution in Table 2.2 is more informative than looking at the raw data in Table 2.1. We see that Papa John's is the most popular restaurant, and that Papa John's is roughly twice as popular as each of the next three runners up—Bruno's, Little Caesars, and Will's. Finally, Pizza Hut and Domino's are the least preferred restaurants.

TABLE 2.1 Pizza Preferences of 50 College Students PizzaPref

Little Caesars	Papa John's	Bruno's	Papa John's	Domino's
Papa John's	Will's Uptown	Papa John's	Pizza Hut	Little Caesars
Pizza Hut	Little Caesars	Will's Uptown	Little Caesars	Bruno's
Papa John's	Bruno's	Papa John's	Will's Uptown	Papa John's
Bruno's	Papa John's	Little Caesars	Papa John's	Little Caesars
Papa John's	Little Caesars	Bruno's	Will's Uptown	Papa John's
Will's Uptown	Papa John's	Will's Uptown	Bruno's	Papa John's
Papa John's	Domino's	Papa John's	Pizza Hut	Will's Uptown
Will's Uptown	Bruno's	Pizza Hut	Papa John's	Papa John's
Little Caesars	Papa John's	Little Caesars	Papa John's	Bruno's

TABLE 2.2 A Frequency Distribution of Pizza Preferences PizzaFreq

Restaurant	Frequency
Bruno's	8
Domino's	2
Little Caesars	9
Papa John's	19
Pizza Hut	4
Will's Uptown	8
	50

TABLE 2.3 Relative Frequency and Percent Frequency Distributions for the Pizza Preference Data PizzaPercents

Restaurant	Relative Frequency	Percent Frequency
Bruno's	8/50 = .16	16%
Domino's	.04	4%
Little Caesars	.18	18%
Papa John's	.38	38%
Pizza Hut	.08	8%
Will's Uptown	.16	16%
	1.0	100%

When we wish to summarize the proportion (or fraction) of items in each class, we employ the **relative frequency** for each class. If the data set consists of n observations, we define the relative frequency of a class as follows:

$$\textbf{Relative frequency of a class} = \frac{\text{frequency of the class}}{n}$$

This quantity is simply the fraction of items in the class. Further, we can obtain the **percent frequency** of a class by multiplying the relative frequency by 100.

Table 2.3 gives a relative frequency distribution and a percent frequency distribution of the pizza preference data. A **relative frequency distribution** is a table that lists the relative frequency for each class, and a **percent frequency distribution** lists the percent frequency for each class. Looking at Table 2.3, we see that the relative frequency for Bruno's pizza is $8/50 = .16$ and that (from the percent frequency distribution) 16% of the sampled students preferred Bruno's pizza. Similarly, the relative frequency for Papa John's pizza is $19/50 = .38$ and 38% of the sampled students preferred Papa John's pizza. Finally, the sum of the relative frequencies in the relative frequency distribution equals 1.0, and the sum of the percent frequencies in the percent frequency distribution equals 100%. These facts are true for all relative frequency and percent frequency distributions.

Part 2: Studying pizza preferences by using bar charts and pie charts A **bar chart** is a graphic that depicts a frequency, relative frequency, or percent frequency distribution. For example, Figure 2.1 gives an Excel bar chart of the pizza preference data. On the horizontal axis we have placed a label for each class (restaurant), while the vertical axis measures frequencies. To construct the bar chart, Excel draws a bar (of fixed width) corresponding to each class label. Each bar is drawn so that its height equals the frequency corresponding to its label. Because the height of each bar is a frequency, we refer to Figure 2.1 as a **frequency bar chart.** Notice that

FIGURE 2.1 Excel Bar Chart of the Pizza Preference Data

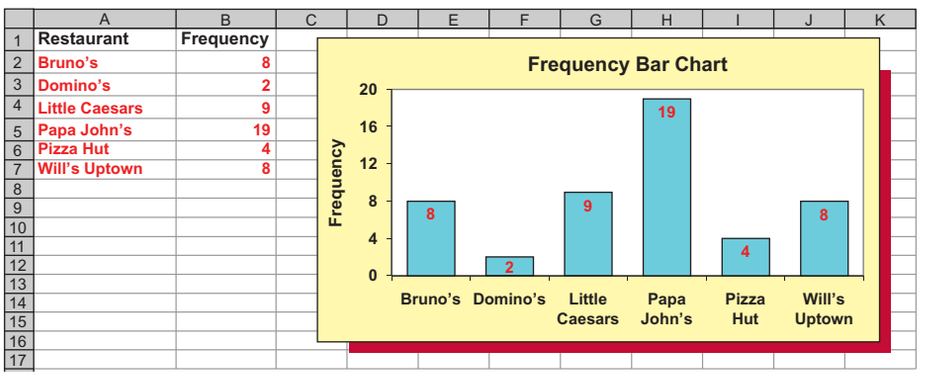
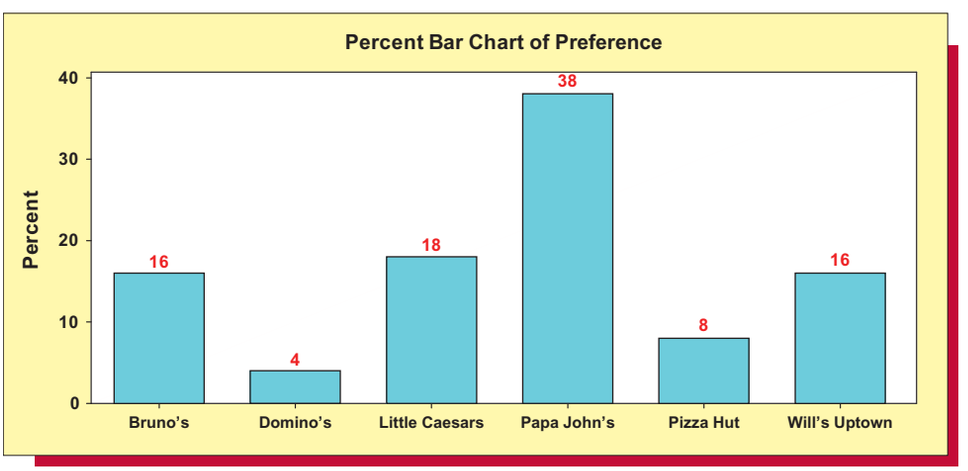


FIGURE 2.2 MINITAB Percent Bar Chart of the Pizza Preference Data



there are gaps between the bars. When data are qualitative, the bars should always be separated by gaps in order to indicate that each class is separate from the others. The bar chart in Figure 2.1 clearly illustrates that, for example, Papa John's pizza is preferred by more sampled students than any other restaurant and Domino's pizza is least preferred by the sampled students.

If desired, the bar heights can represent relative frequencies or percent frequencies. For instance, Figure 2.2 is a MINITAB **percent bar chart** for the pizza preference data. Here the heights of the bars are the percentages given in the percent frequency distribution of Table 2.3. Lastly, the bars in Figures 2.1 and 2.2 have been positioned vertically. Because of this, these bar charts are called **vertical bar charts.** However, sometimes bar charts are constructed with horizontal bars and are called **horizontal bar charts.**

A **pie chart** is another graphic that can be used to depict a frequency distribution. When constructing a pie chart, we first draw a circle to represent the entire data set. We then divide the circle into sectors or "pie slices" based on the relative frequencies of the classes. For example, remembering that a circle consists of 360 degrees, Bruno's Pizza (which has relative frequency .16) is assigned a pie slice that consists of $.16(360) = 57.6$ degrees. Similarly, Papa John's Pizza (with relative frequency .38) is assigned a pie slice having $.38(360) = 136.8$ degrees. The resulting pie chart (constructed using Excel) is shown in Figure 2.3. Here we have labeled the pie slices using the percent frequencies. The pie slices can also be labeled using frequencies or relative frequencies.

FIGURE 2.3 Excel Pie Chart of the Pizza Preference Data

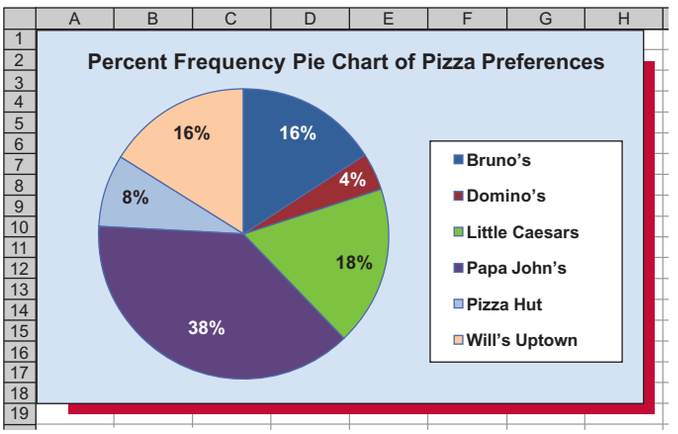
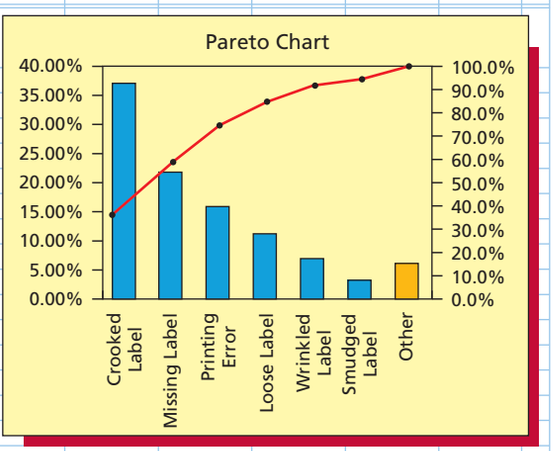


FIGURE 2.4 Excel Frequency Table and Pareto Chart of Labeling Defects DS Labels

	A	B	C	D
1	Frequency Table			
3	Type of Defect	Freq.	Percent	Cum. Percent
4	Crooked Label	78	36.97%	37.0%
5	Missing Label	45	21.33%	58.3%
6	Printing Error	33	15.64%	73.9%
7	Loose Label	23	10.90%	84.8%
8	Wrinkled Label	14	6.64%	91.5%
9	Smudged Label	6	2.84%	94.3%
10	Other	12	5.69%	100.0%
11	*Total*	211	100%	



LO2-2 Construct and interpret Pareto charts (Optional).

The Pareto chart (Optional) **Pareto charts** are used to help identify important quality problems and opportunities for process improvement. By using these charts we can prioritize problem-solving activities. The Pareto chart is named for Vilfredo Pareto (1848–1923), an Italian economist. Pareto suggested that, in many economies, most of the wealth is held by a small minority of the population. It has been found that the **"Pareto principle"** often applies to defects. That is, only a few defect types account for most of a product's quality problems.

To illustrate the use of Pareto charts, suppose that a jelly producer wishes to evaluate the labels being placed on 16-ounce jars of grape jelly. Every day for two weeks, all defective labels found on inspection are classified by type of defect. If a label has more than one defect, the type of defect that is most noticeable is recorded. The Excel output in Figure 2.4 presents the frequencies and percentages of the types of defects observed over the two-week period.

In general, the first step in setting up a **Pareto chart** summarizing data concerning types of defects (or categories) is to construct a frequency table like the one in Figure 2.4. Defects or categories should be listed at the left of the table in *decreasing order by frequencies*—the defect

with the highest frequency will be at the top of the table, the defect with the second-highest frequency below the first, and so forth. If an "other" category is employed, it should be placed at the bottom of the table. The "other" category should not make up 50 percent or more of the total of the frequencies, and the frequency for the "other" category should not exceed the frequency for the defect at the top of the table. If the frequency for the "other" category is too high, data should be collected so that the "other" category can be broken down into new categories. Once the frequency and the percentage for each category are determined, a cumulative percentage for each category is computed. As illustrated in Figure 2.4, the cumulative percentage for a particular category is the sum of the percentages corresponding to the particular category and the categories that are above that category in the table.

A Pareto chart is simply a bar chart having the different kinds of defects or problems listed on the horizontal scale. The heights of the bars on the vertical scale typically represent the frequency of occurrence (or the percentage of occurrence) for each defect or problem. The bars are arranged in decreasing height from left to right. Thus, the most frequent defect will be at the far left, the next most frequent defect to its right, and so forth. If an "other" category is employed, its bar is placed at the far right. The Pareto chart for the labeling defects data is given in Figure 2.4. Here the heights of the bars represent the percentages of occurrences for the different labeling defects, and the vertical scale on the far left corresponds to these percentages. The chart graphically illustrates that crooked labels, missing labels, and printing errors are the most frequent labeling defects.

As is also illustrated in Figure 2.4, a Pareto chart is sometimes augmented by plotting a **cumulative percentage point** for each bar in the Pareto chart. The vertical coordinate of this cumulative percentage point equals the cumulative percentage in the frequency table corresponding to the bar. The cumulative percentage points corresponding to the different bars are connected by line segments, and a vertical scale corresponding to the cumulative percentages is placed on the far right. Examining the cumulative percentage points in Figure 2.4, we see that crooked and missing labels make up 58.3 percent of the labeling defects and that crooked labels, missing labels, and printing errors make up 73.9 percent of the labeling defects.

Technical note The Pareto chart in Figure 2.4 illustrates using an "other" category which combines defect types having low frequencies into a single class. In general, when we employ a frequency distribution, a bar chart, or a pie chart and we encounter classes having small class frequencies, it is common practice to combine the classes into a single "other" category. Classes having frequencies of 5 percent or less are usually handled this way.

Exercises for Section 2.1

CONCEPTS

connect

2.1 Explain the purpose behind constructing a frequency or relative frequency distribution.

2.2 Explain how to compute the relative frequency and percent frequency for each class if you are given a frequency distribution.

2.3 Find an example of a pie chart or bar chart in a newspaper or magazine. Copy it, and hand it in with a written analysis of the information conveyed by the chart.

METHODS AND APPLICATIONS

2.4 A multiple choice question on an exam has four possible responses—(a), (b), (c), and (d). When 250 students take the exam, 100 give response (a), 25 give response (b), 75 give response (c), and 50 give response (d).

- **a** Write out the frequency distribution, relative frequency distribution, and percent frequency distribution for these responses.
- **b** Construct a bar chart for these data using frequencies.

2.5 Consider constructing a pie chart for the exam question responses in Exercise 2.4.

- **a** How many degrees (out of 360) would be assigned to the "pie slice" for the response (a)?
- **b** How many degrees would be assigned to the "pie slice" for response (b)?
- **c** Construct the pie chart for the exam question responses.

Product	Relative Frequency
W	.15
X	—
Y	.36
Z	.28

2.6 Consider the partial relative frequency distribution of consumer preferences for four products—W, X, Y, and Z—that is shown in the page margin.

a Find the relative frequency for product X.
b If 500 consumers were surveyed, give the frequency distribution for these data.
c Construct a percent frequency bar chart for these data.
d If we wish to depict these data using a pie chart, find how many degrees (out of 360) should be assigned to each of products W, X, Y, and Z. Then construct the pie chart.

2.7 Below we give the overall dining experience ratings (Outstanding, Very Good, Good, Average, or Poor) of 30 randomly selected patrons at a restaurant on a Saturday evening. DS RestRating

Outstanding	Good	Very Good	Very Good	Outstanding	Good
Outstanding	Outstanding	Outstanding	Very Good	Very Good	Average
Very Good	Outstanding	Outstanding	Outstanding	Outstanding	Very Good
Outstanding	Good	Very Good	Outstanding	Very Good	Outstanding
Good	Very Good	Outstanding	Very Good	Good	Outstanding

a Find the frequency distribution and relative frequency distribution for these data.
b Construct a percentage bar chart for these data.
c Construct a percentage pie chart for these data.

2.8 Fifty randomly selected adults who follow professional sports were asked to name their favorite professional sports league. The results are as follows where MLB = Major League Baseball, MLS = Major League Soccer, NBA = National Basketball Association, NFL = National Football League, and NHL = National Hockey League. DS ProfSports

NFL	NBA	NFL	MLB	MLB	NHL	NFL	NFL	MLS	MLB
MLB	NFL	MLB	NBA	NBA	NFL	NFL	NFL	NHL	NBA
NBA	NFL	NHL	NFL	MLS	NFL	MLB	NFL	MLB	NFL
NHL	MLB	NHL	NFL	NFL	NFL	MLB	NFL	NBA	NFL
MLS	NFL	MLB	NBA	NFL	NFL	MLB	NBA	NFL	NFL

a Find the frequency distribution, relative frequency distribution, and percent frequency distribution for these data.
b Construct a frequency bar chart for these data.
c Construct a pie chart for these data.
d Which professional sports league is most popular with these 50 adults? Which is least popular?

2.9 The National Automobile Dealers Association (NADA) publishes *AutoExec* magazine, which annually reports on new vehicle sales and market shares by manufacturer. As given on the *AutoExec* magazine website in May 2006, new vehicle market shares in the United States for 2005 were as follows[1]: Chrysler/Dodge/Jeep 13.6%, Ford 18.3%, GM 26.3%, Japanese (Toyota/Honda/Nissan) 28.3%, other imports 13.5%. Construct a percent frequency bar chart and a percentage pie chart for the 2005 auto market shares. DS AutoShares05

2.10 Figure 2.5 gives a percentage pie chart of new vehicle market shares in the U.S. for 2010 as given by GoodCarBadCar.net. Use this pie chart and your results from Exercise 2.9 to write an analysis explaining how new vehicle market shares in the United States have changed from 2005 to 2010. DS AutoShares10

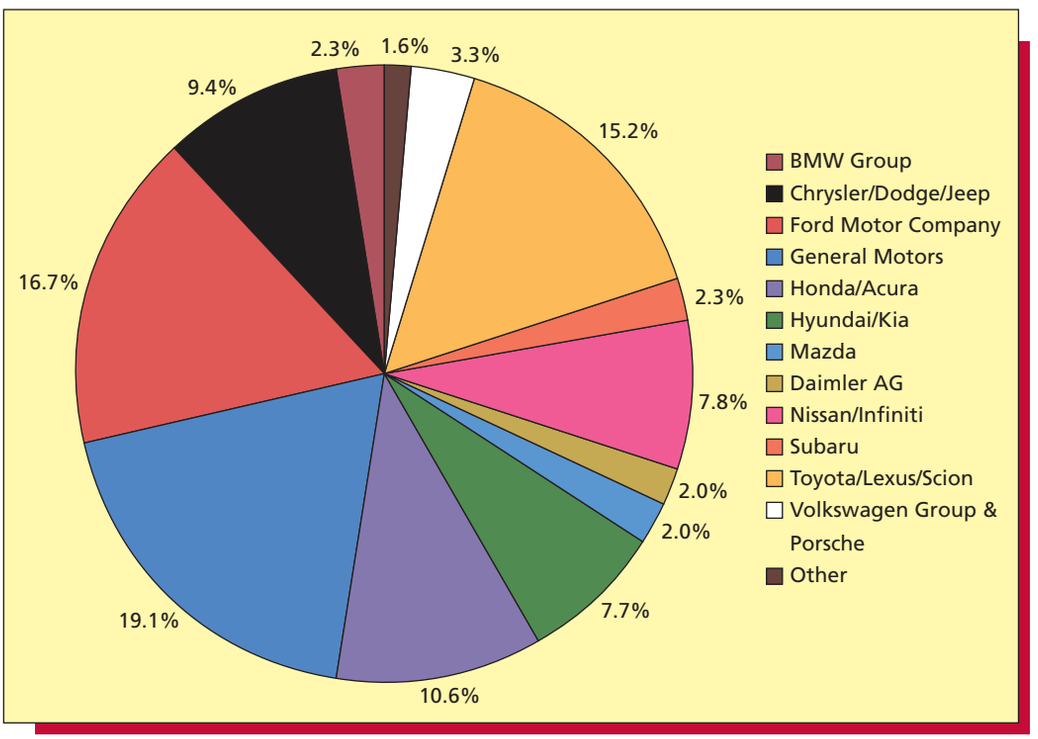
2.11 On January 11, 2005, the Gallup Organization released the results of a poll investigating how many Americans have private health insurance. The results showed that among Americans making less than $30,000 per year, 33% had private insurance, 50% were covered by Medicare/Medicaid, and 17% had no health insurance, while among Americans making $75,000 or more per year, 87% had private insurance, 9% were covered by Medicare/Medicaid, and 4% had no health insurance.[2] Use bar and pie charts to compare health coverage of the two income groups.

2.12 In an article in *Quality Progress,* Barbara A. Cleary reports on improvements made in a software supplier's responses to customer calls. In this article, the author states:

> In an effort to improve its response time for these important customer-support calls, an inbound telephone inquiry team was formed at PQ Systems, Inc., a software and training organization in Dayton, Ohio. The team found that 88 percent of the customers' calls were already being answered immediately by the technical support group, but those who had to be called back had to wait an average of 56.6 minutes. No customer complaints had been registered, but the team believed that this response rate could be improved.

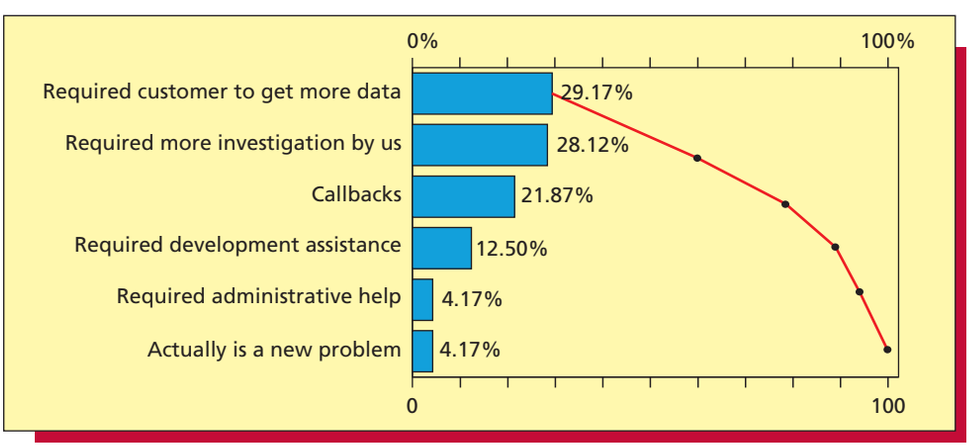
[1] **Source:** www.autoexecmag.com, May 15, 2006.
[2] **Source:** http://gallup.com/poll/content/default.aspx?ci=14581.

FIGURE 2.5 **A Pie Chart of U.S. Automobile Sales in 2010 as given by GoodCarBadCar.net (for Exercise 2.10)**
AutoShares10



Source: http://www.goodcarbadcar.net/2011/01/new-vehicle-market-share-by-brand-in_05.html (accessed 5/29/12).

FIGURE 2.6 **A Pareto Chart for Incomplete Customer Calls (for Exercise 2.12)**



Source: B. A. Cleary, "Company Cares about Customers' Calls," *Quality Progress* (November 1993), pp. 60–73. Copyright © 1993 American Society for Quality Control. Used with permission.

As part of its improvement process, the company studied the disposition of complete and incomplete calls to its technical support analysts. A call is considered complete if the customer's problem has been resolved; otherwise the call is incomplete. Figure 2.6 shows a Pareto chart analysis for the incomplete customer calls.

a What percentage of incomplete calls required "more investigation" by the analyst or "administrative help"?

b What percentage of incomplete calls actually presented a "new problem"?

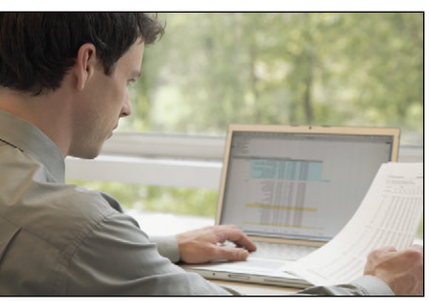
c In light of your answers to *a* and *b*, can you make a suggestion?

LO2-3 Summarize quantitative data by using frequency distributions, histograms, frequency polygons, and ogives.

2.2 Graphically Summarizing Quantitative Data

Frequency distributions and histograms We often need to summarize and describe the shape of the distribution of a population or sample of **measurements**. Such data are often summarized by grouping the measurements into the classes of a frequency distribution and by displaying the data in the form of a **histogram.** We explain how to construct a histogram in the following example.

EXAMPLE 2.2 The e-billing Case: Reducing Bill Payment Times[3]



Major consulting firms such as Accenture, Ernst & Young Consulting, and Deloitte & Touche Consulting employ statistical analysis to assess the effectiveness of the systems they design for their customers. In this case a consulting firm has developed an electronic billing system for a Hamilton, Ohio, trucking company. The system sends invoices electronically to each customer's computer and allows customers to easily check and correct errors. It is hoped that the new billing system will substantially reduce the amount of time it takes customers to make payments. Typical payment times—measured from the date on an invoice to the date payment is received—using the trucking company's old billing system had been 39 days or more. This exceeded the industry standard payment time of 30 days.

The new billing system does not automatically compute the payment time for each invoice because there is no continuing need for this information. Therefore, in order to assess the system's effectiveness, the consulting firm selects a random sample of 65 invoices from the 7,823 invoices processed during the first three months of the new system's operation. The payment times for the 65 sample invoices are manually determined and are given in Table 2.4. If this sample can be used to establish that the new billing system substantially reduces payment times, the consulting firm plans to market the system to other trucking firms.

Looking at the payment times in Table 2.4, we can see that the shortest payment time is 10 days and that the longest payment time is 29 days. Beyond that, it is pretty difficult to interpret the data in any meaningful way. To better understand the sample of 65 payment times, the consulting firm will form a frequency distribution of the data and will graph the distribution by constructing a histogram. Similar to the frequency distributions for qualitative data we studied in Section 2.1, the frequency distribution will divide the payment times into classes and will tell us how many of the payment times are in each class.

Step 1: Find the number of classes One rule for finding an appropriate number of classes says that the number of classes should be the smallest whole number K that makes the quantity 2^K greater than the number of measurements in the data set. For the payment time data we have 65 measurements. Because $2^6 = 64$ is less than 65 and $2^7 = 128$ is greater than 65, we should use $K = 7$ classes. Table 2.5 gives the appropriate number of classes (determined by the 2^K rule) to use for data sets of various sizes.

Step 2: Find the class length We find the length of each class by computing

$$\text{approximate class length} = \frac{\text{largest measurement} - \text{smallest measurement}}{\text{number of classes}}$$

TABLE 2.4 A Sample of Payment Times (in Days) for 65 Randomly Selected Invoices PayTime

22	29	16	15	18	17	12	13	17	16	15
19	17	10	21	15	14	17	18	12	20	14
16	15	16	20	22	14	25	19	23	15	19
18	23	22	16	16	19	13	18	24	24	26
13	18	17	15	24	15	17	14	18	17	21
16	21	25	19	20	27	16	17	16	21	

[3]This case is based on a real problem encountered by a company that employs one of our former students. For purposes of confidentiality, we have withheld the company's name.

TABLE 2.5 Recommended Number of Classes for Data Sets of n Measurements*

Number of Classes	Size, n, of the Data Set
2	$1 \le n < 4$
3	$4 \le n < 8$
4	$8 \le n < 16$
5	$16 \le n < 32$
6	$32 \le n < 64$
7	$64 \le n < 128$
8	$128 \le n < 256$
9	$256 \le n < 528$
10	$528 \le n < 1056$

*For completeness sake we have included all values of $n \ge 1$ in this table. However, we do not recommend constructing a histogram with fewer than 16 measurements.

TABLE 2.6 Seven Nonoverlapping Classes for a Frequency Distribution of the 65 Payment Times

Class 1	10 days and less than 13 days
Class 2	13 days and less than 16 days
Class 3	16 days and less than 19 days
Class 4	19 days and less than 22 days
Class 5	22 days and less than 25 days
Class 6	25 days and less than 28 days
Class 7	28 days and less than 31 days

Because the largest and smallest payment times in Table 2.4 are 29 days and 10 days, the approximate class length is $(29 - 10)/7 = 2.7143$. To obtain an easier to read final class length, we round this value. Commonly, the approximate class length is rounded up to the precision of the data measurements (that is, increased to the next number that has the same number of decimal places as the data measurements). For instance, because the payment times are measured to the nearest day, we round 2.7143 days up to 3 days.

Step 3: Form nonoverlapping classes of equal width We can form the classes of the frequency distribution by defining the **boundaries** of the classes. To find the first class boundary, we find the smallest payment time in Table 2.4, which is 10 days. This value is the lower boundary of the first class. Adding the class length of 3 to this lower boundary, we obtain $10 + 3 = 13$, which is the upper boundary of the first class and the lower boundary of the second class. Similarly, the upper boundary of the second class and the lower boundary of the third class equals $13 + 3 = 16$. Continuing in this fashion, the lower boundaries of the remaining classes are 19, 22, 25, and 28. Adding the class length 3 to the lower boundary of the last class gives us the upper boundary of the last class, 31. These boundaries define seven nonoverlapping classes for the frequency distribution. We summarize these classes in Table 2.6. For instance, the first class—10 days and less than 13 days—includes the payment times 10, 11, and 12 days; the second class—13 days and less than 16 days—includes the payment times 13, 14, and 15 days; and so forth. Notice that the largest *observed* payment time—29 days—is contained in the last class. In cases where the largest measurement is not contained in the last class, we simply add another class. Generally speaking, the guidelines we have given for forming classes are not inflexible rules. Rather, they are intended to help us find reasonable classes. Finally, the method we have used for forming classes results in classes of equal length. Generally, forming classes of equal length will make it easier to appropriately interpret the frequency distribution.

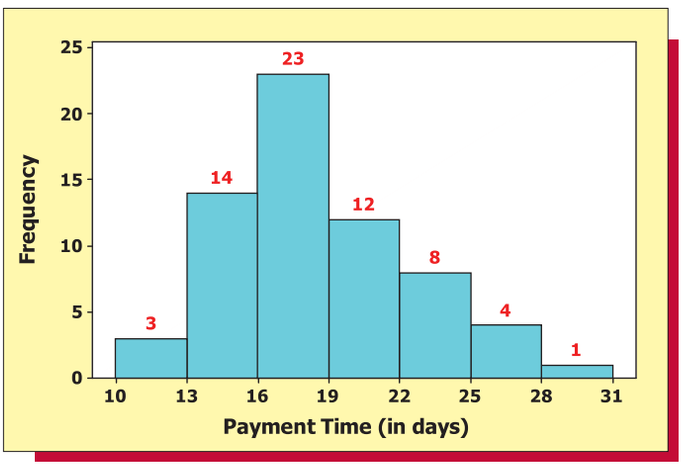
Step 4: Tally and count the number of measurements in each class Having formed the classes, we now count the number of measurements that fall into each class. To do this, it is convenient to tally the measurements. We simply list the classes, examine the payment times in Table 2.4 one at a time, and record a tally mark corresponding to a particular class each time we encounter a measurement that falls in that class. For example, because the first four payment times in Table 2.4 are 22, 19, 16, and 18, the first four tally marks are shown below. Here, for brevity, we express the class "10 days and less than 13 days" as "10 < 13" and use similar notation for the other classes.

Class	First 4 Tally Marks	All 65 Tally Marks	Frequency
10 < 13		///	3
13 < 16		𝍸 𝍸 ////	14
16 < 19	//	𝍸 𝍸 𝍸 𝍸 ///	23
19 < 22	/	𝍸 𝍸 //	12
22 < 25	/	𝍸 ///	8
25 < 28		////	4
28 < 31		/	1

TABLE 2.7 Frequency Distributions of the 65 Payment Times

Class	Frequency	Relative Frequency	Percent Frequency
10 < 13	3	3/65 = .0462	4.62%
13 < 16	14	14/65 = .2154	21.54
16 < 19	23	.3538	35.38
19 < 22	12	.1846	18.46
22 < 25	8	.1231	12.31
25 < 28	4	.0615	6.15
28 < 31	1	.0154	1.54

FIGURE 2.7 A Frequency Histogram of the 65 Payment Times



After examining all 65 payment times, we have recorded 65 tally marks—see the bottom of page 43. We find the **frequency** for each class by counting the number of tally marks recorded for the class. For instance, counting the number of tally marks for the class "13 < 16", we obtain the frequency 14 for this class. The frequencies for all seven classes are summarized in Table 2.7. This summary is the **frequency distribution** for the 65 payment times. Table 2.7 also gives the *relative frequency* and the *percent frequency* for each of the seven classes. The **relative frequency** of a class is the proportion (fraction) of the total number of measurements that are in the class. For example, there are 14 payment times in the second class, so its relative frequency is $14/65 = .2154$. This says that the proportion of the 65 payment times that are in the second class is .2154, or, equivalently, that $100(.2154)\% = 21.54\%$ of the payment times are in the second class. A list of all of the classes—along with each class relative frequency—is called a **relative frequency distribution.** A list of all of the classes—along with each class percent frequency—is called a **percent frequency distribution.**

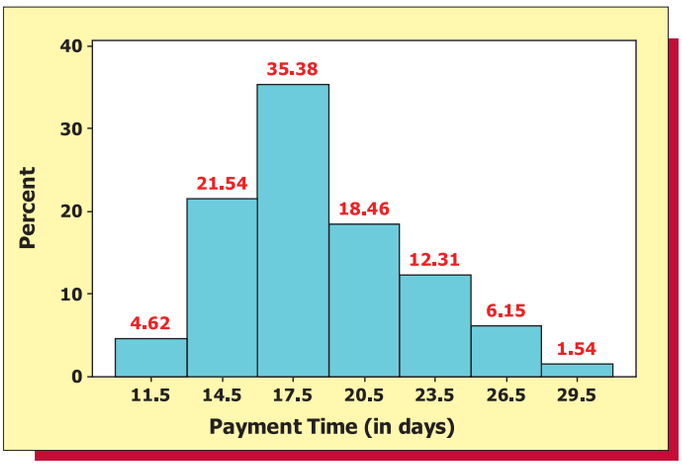
Step 5: Graph the histogram We can graphically portray the distribution of payment times by drawing a **histogram.** The histogram can be constructed using the frequency, relative frequency, or percent frequency distribution. To set up the histogram, we draw rectangles that correspond to the classes. The base of the rectangle corresponding to a class represents the payment times in the class. The height of the rectangle can represent the class frequency, relative frequency, or percent frequency.

We have drawn a **frequency histogram** of the 65 payment times in Figure 2.7. The first (leftmost) rectangle, or "bar," of the histogram represents the payment times 10, 11, and 12. Looking at Figure 2.7, we see that the base of this rectangle is drawn from the lower boundary (10) of the first class in the frequency distribution of payment times to the lower boundary (13) of the second class. The height of this rectangle tells us that the frequency of the first class is 3. The second histogram rectangle represents payment times 13, 14, and 15. Its base is drawn from the lower boundary (13) of the second class to the lower boundary (16) of the third class, and its height tells us that the frequency of the second class is 14. The other histogram bars are constructed similarly. Notice that there are no gaps between the adjacent rectangles in the histogram. Here, although the payment times have been recorded to the nearest whole day, the fact that the histogram bars touch each other emphasizes that a payment time could (in theory) be any number on the horizontal axis. In general, histograms are drawn so that adjacent bars touch each other.

Looking at the frequency distribution in Table 2.7 and the frequency histogram in Figure 2.7, we can describe the payment times:

1. None of the payment times exceeds the industry standard of 30 days. (Actually, all of the payment times are less than 30—remember the largest payment time is 29 days.)
2. The payment times are concentrated between 13 and 24 days (57 of the 65, or $(57/65) \times 100 = 87.69\%$, of the payment times are in this range).
3. More payment times are in the class "16 < 19" than are in any other class (23 payment times are in this class).

FIGURE 2.8 A Percent Frequency Histogram of the 65 Payment Times



Notice that the frequency distribution and histogram allow us to make some helpful conclusions about the payment times, whereas looking at the raw data (the payment times in Table 2.4 on page 42) did not.

A **relative frequency histogram** and a **percent frequency histogram** of the payment times would both be drawn like Figure 2.7 except that the heights of the rectangles would represent, respectively, the relative frequencies and the percent frequencies in Table 2.7. For example, Figure 2.8 gives a percent frequency histogram of the payment times. This histogram also illustrates that we sometimes label the classes on the horizontal axis using the **class midpoints.** Each class midpoint is exactly halfway between the boundaries of its class. For instance, the midpoint of the first class, 11.5, is halfway between the class boundaries 10 and 13. The midpoint of the second class, 14.5, is halfway between the class boundaries 13 and 16. The other class midpoints are found similarly. The percent frequency distribution of Figure 2.8 tells us that 21.54% of the payment times are in the second class (which has midpoint 14.5 and represents the payment times 13, 14, and 15).

In the following box we summarize the steps needed to set up a frequency distribution and histogram:

Constructing Frequency Distributions and Histograms

1 Find the number of classes. Generally, the number of classes K should equal the smallest whole number that makes the quantity 2^K greater than the total number of measurements n (see Table 2.5 on page 43).

2 Compute the approximate **class length:**

$$\frac{\text{largest measurement} - \text{smallest measurement}}{K}$$

Often the final class length is obtained by rounding this value up to the same level of precision as the data.

3 Form nonoverlapping classes of equal length. Form the classes by finding the **class boundaries.** The lower boundary of the first class is the smallest measurement in the data set. Add the class length to this boundary to obtain the next boundary. Successive boundaries are found by repeatedly adding the class length until the upper boundary of the last (Kth) class is found.

4 Tally and count the number of measurements in each class. The **frequency** for each class is the count of the number of measurements in the class. The **relative frequency** for each class is the fraction of measurements in the class. The **percent frequency** for each class is its relative frequency multiplied by 100%.

5 Graph the histogram. To draw a **frequency histogram,** plot each frequency as the height of a rectangle positioned over its corresponding class. Use the class boundaries to separate adjacent rectangles. A **relative frequency histogram** and a **percent histogram** are graphed in the same way except that the heights of the rectangles are, respectively, the relative frequencies and the percent frequencies.

The procedure in the above box is not the only way to construct a histogram. Often, histograms are constructed more informally. For instance, it is not necessary to set the lower boundary of the first (leftmost) class equal to the smallest measurement in the data. As an example, suppose that we wish to form a histogram of the 50 gas mileages given in Table 1.6 (page 11). Examining the mileages, we see that the smallest mileage is 29.8 mpg and that the largest mileage is 33.3 mpg. Therefore, it would be convenient to begin the first (leftmost) class at 29.5 mpg and end the last (rightmost) class at 33.5 mpg. Further, it would be reasonable to use classes that are .5 mpg in length. We would then use 8 classes: $29.5 < 30$, $30 < 30.5$, $30.5 < 31$, $31 < 31.5$, $31.5 < 32$, $32 < 32.5$, $32.5 < 33$, and $33 < 33.5$. A histogram of the gas mileages employing these classes is shown in Figure 2.9.

Sometimes it is desirable to let the nature of the problem determine the histogram classes. For example, to construct a histogram describing the ages of the residents in a city, it might be reasonable to use classes having 10-year lengths (that is, under 10 years, 10–19 years, 20–29 years, 30–39 years, and so on).

Notice that in our examples we have used classes having equal class lengths. In general, it is best to use equal class lengths whenever the raw data (that is, all the actual measurements) are available. However, sometimes histograms are formed with unequal class lengths—particularly when we are using published data as a source. Economic data and data in the social sciences are often published in the form of frequency distributions having unequal class lengths. Dealing with this kind of data is discussed in Exercise 2.81. Also discussed in this exercise is how to deal with **open-ended** classes. For example, if we are constructing a histogram describing the yearly incomes of U.S. households, an open-ended class could be households earning over $500,000 per year.

As an alternative to constructing a frequency distribution and histogram by hand, we can use software packages such as Excel and MINITAB. Each of these packages will automatically define histogram classes for the user. However, these automatically defined classes will not necessarily be the same as those that would be obtained using the manual method we have previously described. Furthermore, the packages define classes by using different methods. (Descriptions of how the classes are defined can often be found in help menus.) For example, Figure 2.10 gives a MINITAB frequency histogram of the payment times in Table 2.4. Here, MINITAB has defined 11 classes and has labeled five of the classes on the horizontal axis using midpoints (12, 16, 20, 24, 28). It is easy to see that the midpoints of the unlabeled classes are 10, 14, 18, 22, 26, and 30. Moreover, the boundaries of the first class are 9 and 11, the boundaries of the second class are 11 and 13, and so forth. MINITAB counts frequencies as we have previously described. For instance, one payment time is at least 9 and less than 11, two payment times are at least 11 and less than 13, seven payment times are at least 13 and less than 15, and so forth.

FIGURE 2.9 **A Percent Frequency Histogram of the Gas Mileages: The Gas Mileage Distribution Is Symmetrical and Mound Shaped**

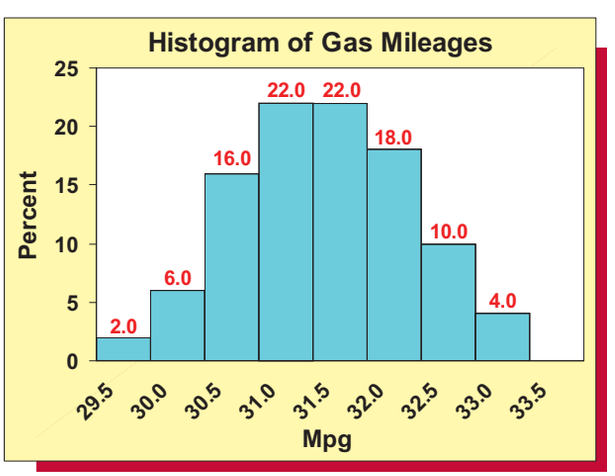


FIGURE 2.10 **A MINITAB Frequency Histogram of the Payment Times with Automatic Classes: The Payment Time Distribution Is Skewed to the Right**

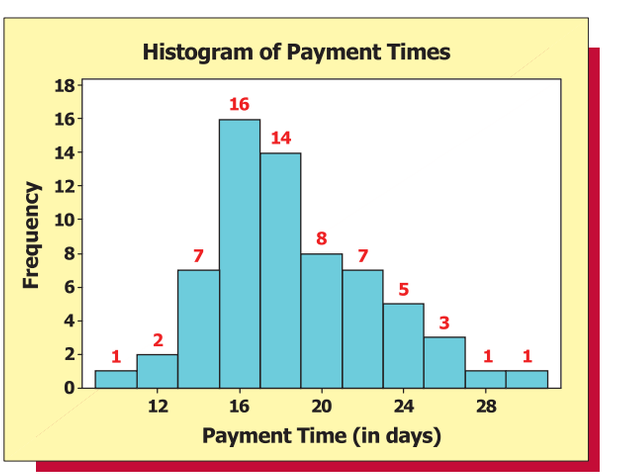


FIGURE 2.11 An Excel Frequency Histogram of the Bottle Design Ratings: The Distribution of Ratings Is Skewed to the Left

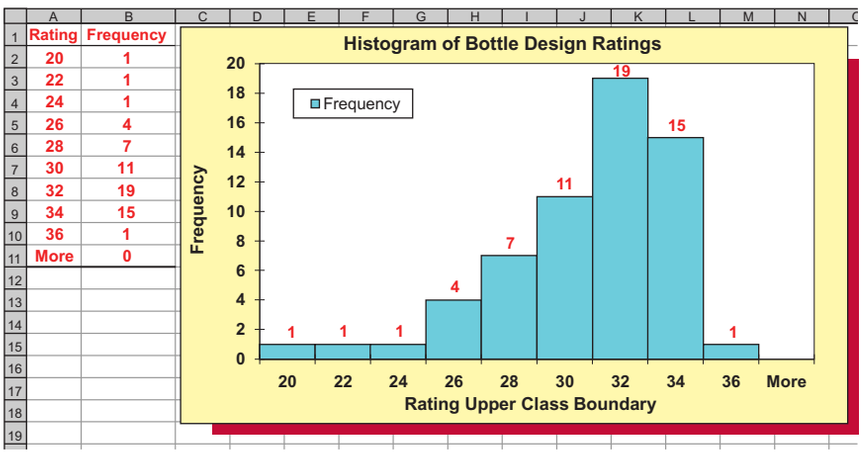


Figure 2.11 gives an Excel frequency distribution and histogram of the bottle design ratings in Table 1.5. Excel labels histogram classes using their upper class boundaries. For example, the first class has an upper class boundary equal to the smallest rating of 20 and contains only this smallest rating. The boundaries of the second class are 20 and 22, the boundaries of the third class are 22 and 24, and so forth. The last class corresponds to ratings more than 36. Excel's method for counting frequencies differs from that of MINITAB (and, therefore, also differs from the way we counted frequencies by hand in Example 2.2). Excel assigns a frequency to a particular class by counting the number of measurements that are greater than the lower boundary of the class and less than or equal to the upper boundary of the class. For example, one bottle design rating is greater than 20 and less than or equal to (that is, at most) 22. Similarly, 15 bottle design ratings are greater than 32 and at most 34.

In Figure 2.10 we have used MINITAB to automatically form histogram classes. It is also possible to force software packages to form histogram classes that are defined by the user. We explain how to do this in the appendices at the end of this chapter. Because Excel does not always automatically define acceptable classes, the classes in Figure 2.11 are a modification of Excel's automatic classes. We also explain this modification in the appendices at the end of this chapter.

Some common distribution shapes We often graph a frequency distribution in the form of a histogram in order to visualize the *shape* of the distribution. If we look at the histogram of payment times in Figure 2.10, we see that the right tail of the histogram is longer than the left tail. When a histogram has this general shape, we say that the distribution is **skewed to the right.** Here the long right tail tells us that a few of the payment times are somewhat longer than the rest. If we look at the histogram of bottle design ratings in Figure 2.11, we see that the left tail of the histogram is much longer than the right tail. When a histogram has this general shape, we say that the distribution is **skewed to the left.** Here the long tail to the left tells us that, while most of the bottle design ratings are concentrated above 25 or so, a few of the ratings are lower than the rest. Finally, looking at the histogram of gas mileages in Figure 2.9, we see that the right and left tails of the histogram appear to be mirror images of each other. When a histogram has this general shape, we say that the distribution is **symmetrical.** Moreover, the distribution of gas mileages appears to be piled up in the middle or **mound shaped.**

Mound-shaped, symmetrical distributions as well as distributions that are skewed to the right or left are commonly found in practice. For example, distributions of scores on standardized tests such as the SAT and ACT tend to be mound shaped and symmetrical, whereas distributions of scores on tests in college statistics courses might be skewed to the left—a few students don't study and get scores much lower than the rest. On the other hand, economic data such as income data are often skewed to the right—a few people have incomes much higher than most others.

Many other distribution shapes are possible. For example, some distributions have two or more peaks—we will give an example of this distribution shape later in this section. It is often very useful to know the shape of a distribution. For example, knowing that the distribution of bottle design ratings is skewed to the left suggests that a few consumers may have noticed a problem with design that others didn't see. Further investigation into why these consumers gave the design low ratings might allow the company to improve the design.

Frequency polygons Another graphical display that can be used to depict a frequency distribution is a **frequency polygon.** To construct this graphic, we plot a point above each class midpoint at a height equal to the frequency of the class—the height can also be the class relative frequency or class percent frequency if so desired. Then we connect the points with line segments. As we will demonstrate in the following example, this kind of graphic can be particularly useful when we wish to compare two or more distributions.

EXAMPLE 2.3 Comparing Two Grade Distributions

Table 2.8 lists (in increasing order) the scores earned on the first exam by the 40 students in a business statistics course taught by one of the authors several semesters ago. Figure 2.12 gives a percent frequency polygon for these exam scores. Because exam scores are often reported by using 10-point grade ranges (for instance, 80 to 90 percent), we have defined the following classes: $30 < 40$, $40 < 50$, $50 < 60$, $60 < 70$, $70 < 80$, $80 < 90$, and $90 < 100$. This is an example of letting the situation determine the classes of a frequency distribution, which is common practice when the situation naturally defines classes. The points that form the polygon have been plotted corresponding to the midpoints of the classes (35, 45, 55, 65, 75, 85, 95). Each point is plotted at a height that equals the percentage of exam scores in its class. For instance, because 10 of the 40 scores are at least 90 and less than 100, the plot point corresponding to the class midpoint 95 is plotted at a height of 25 percent.

Looking at Figure 2.12, we see that there is a concentration of scores in the 85 to 95 range and another concentration of scores around 65. In addition, the distribution of scores is somewhat skewed to the left—a few students had scores (in the 30s and 40s) that were quite a bit lower than the rest.

This is an example of a distribution having two peaks. When a distribution has multiple peaks, finding the reason for the different peaks often provides useful information. The reason for the two-peaked distribution of exam scores was that some students were not attending class regularly. Students who received scores in the 60s and below admitted that they were cutting class, whereas students who received higher scores were attending class on a regular basis.

After identifying the reason for the concentration of lower scores, the instructor established an attendance policy that forced students to attend every class—any student who missed a class was

TABLE 2.8 Exam Scores for the First Exam Given in a Statistics Class DS FirstExam

32	63	69	85	91
45	64	69	86	92
50	64	72	87	92
56	65	76	87	93
58	66	78	88	93
60	67	81	89	94
61	67	83	90	96
61	68	83	90	98

FIGURE 2.12 A Percent Frequency Polygon of the Exam Scores

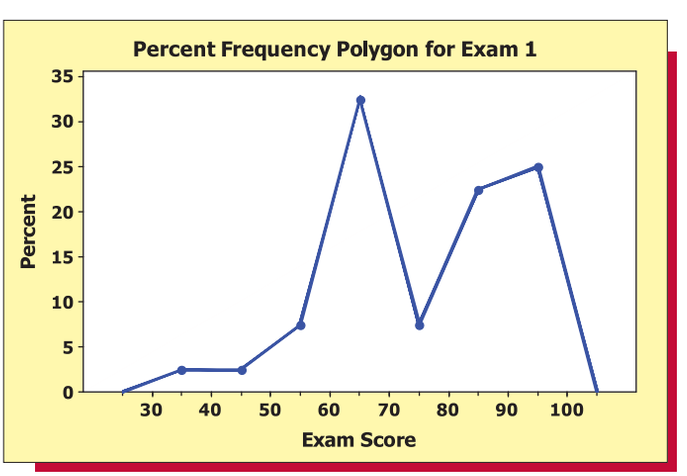
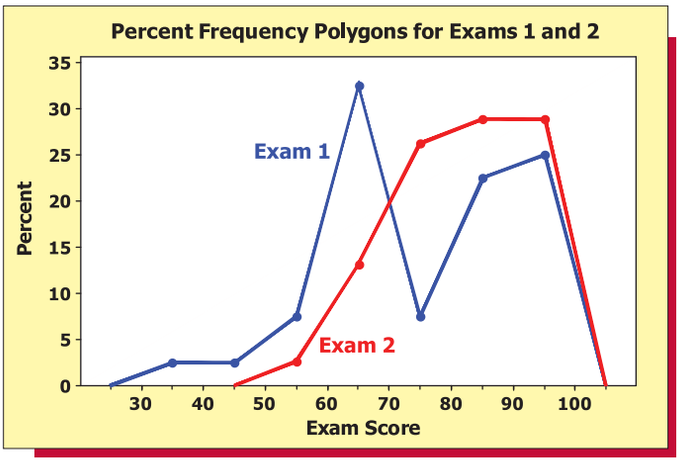


TABLE 2.9 Exam Scores for the Second Statistics Exam—after a New Attendance Policy DS SecondExam

55	74	80	87	93
62	74	82	88	94
63	74	83	89	94
66	75	84	90	95
67	76	85	91	97
67	77	86	91	99
71	77	86	92	
73	78	87	93	

FIGURE 2.13 Percent Frequency Polygons of the Scores on the First Two Exams in a Statistics Course



to be dropped from the course. Table 2.9 presents the scores on the second exam—after the new attendance policy. Figure 2.13 presents (and allows us to compare) the percent frequency polygons for both exams. We see that the polygon for the second exam is single peaked—the attendance policy[4] eliminated the concentration of scores in the 60s, although the scores are still somewhat skewed to the left.

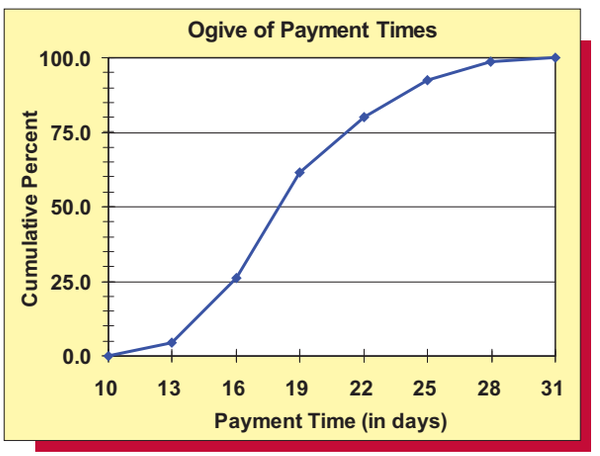
Cumulative distributions and ogives Another way to summarize a distribution is to construct a **cumulative distribution.** To do this, we use the same number of classes, the same class lengths, and the same class boundaries that we have used for the frequency distribution of a data set. However, in order to construct a **cumulative frequency distribution,** we record for each class *the number of measurements that are less than the upper boundary of the class.* To illustrate this idea, Table 2.10 gives the cumulative frequency distribution of the payment time distribution summarized in Table 2.7 (page 44). Columns (1) and (2) in this table give the frequency distribution of the payment times. Column (3) gives the **cumulative frequency** for each class. To see how these values are obtained, the cumulative frequency for the class $10 < 13$ is the number of payment times less than 13. This is obviously the frequency for the class $10 < 13$, which is 3. The cumulative frequency for the class $13 < 16$ is the number of payment times less than 16, which is obtained by adding the frequencies for the first two classes—that is, $3 + 14 = 17$. The cumulative frequency for the class $16 < 19$ is the number of payment times less than 19—that is, $3 + 14 + 23 = 40$. We see that, in general, a cumulative frequency is obtained by summing the frequencies of all classes representing values less than the upper boundary of the class.

TABLE 2.10 A Frequency Distribution, Cumulative Frequency Distribution, Cumulative Relative Frequency Distribution, and Cumulative Percent Frequency Distribution for the Payment Time Data

(1) Class	(2) Frequency	(3) Cumulative Frequency	(4) Cumulative Relative Frequency	(5) Cumulative Percent Frequency
$10 < 13$	3	3	3/65 = .0462	4.62%
$13 < 16$	14	17	17/65 = .2615	26.15
$16 < 19$	23	40	.6154	61.54
$19 < 22$	12	52	.8000	80.00
$22 < 25$	8	60	.9231	92.31
$25 < 28$	4	64	.9846	98.46
$28 < 31$	1	65	1.0000	100.00

[4]Other explanations are possible. For instance, all of the students who did poorly on the first exam might have studied harder for the second exam. However, the instructor's 30 years of teaching experience suggest that attendance was the critical factor.

FIGURE 2.14 A Percent Frequency Ogive of the Payment Times



Column (4) gives the **cumulative relative frequency** for each class, which is obtained by summing the relative frequencies of all classes representing values less than the upper boundary of the class. Or, more simply, this value can be found by dividing the cumulative frequency for the class by the total number of measurements in the data set. For instance, the cumulative relative frequency for the class $19 < 22$ is $52/65 = .8$. Column (5) gives the **cumulative percent frequency** for each class, which is obtained by summing the percent frequencies of all classes representing values less than the upper boundary of the class. More simply, this value can be found by multiplying the cumulative relative frequency of a class by 100. For instance, the cumulative percent frequency for the class $19 < 22$ is .8 (100) = 80 percent.

As an example of interpreting Table 2.10, 60 of the 65 payment times are 24 days or less, or, equivalently, 92.31 percent of the payment times (or a fraction of .9231 of the payment times) are 24 days or less. Also, notice that the last entry in the cumulative frequency distribution is the total number of measurements (here, 65 payment times). In addition, the last entry in the cumulative relative frequency distribution is 1.0 and the last entry in the cumulative percent frequency distribution is 100%. In general, for any data set, these last entries will be, respectively, the total number of measurements, 1.0, and 100%.

An **ogive** (pronounced "oh-jive") is a graph of a cumulative distribution. To construct a frequency ogive, we plot a point above each upper class boundary at a height equal to the cumulative frequency of the class. We then connect the plotted points with line segments. A similar graph can be drawn using the cumulative relative frequencies or the cumulative percent frequencies. As an example, Figure 2.14 gives a percent frequency ogive of the payment times. Looking at this figure, we see that, for instance, a little more than 25 percent (actually, 26.15 percent according to Table 2.10) of the payment times are less than 16 days, while 80 percent of the payment times are less than 22 days. Also notice that we have completed the ogive by plotting an additional point at the lower boundary of the first (leftmost) class at a height equal to zero. This depicts the fact that none of the payment times is less than 10 days. Finally, the ogive graphically shows that all (100 percent) of the payment times are less than 31 days.

Exercises for Section 2.2

CONCEPTS

connect

2.13 Explain:

- **a** Why we construct a frequency distribution and a histogram for a data set.
- **b** The difference between a frequency histogram and a frequency polygon.
- **c** The difference between a frequency polygon and a frequency ogive.

2.14 Explain how to find:

- **a** The frequency for a class.
- **b** The relative frequency for a class.
- **c** The percent frequency for a class.

2.15 Explain what each of the following distribution shapes looks like. Then draw a picture that illustrates each shape.

- **a** Symmetrical and mound shaped
- **b** Double peaked
- **c** Skewed to the right
- **d** Skewed to the left

METHODS AND APPLICATIONS

2.16 Consider the following data: DS HistoData

36	39	36	35	36	20	19
46	40	42	34	41	36	42
40	38	33	37	22	33	28
38	38	34	37	17	25	38

- **a** Find the number of classes needed to construct a histogram.
- **b** Find the class length.
- **c** Define nonoverlapping classes for a frequency distribution.
- **d** Tally the number of values in each class and develop a frequency distribution.
- **e** Draw a histogram for these data.
- **f** Develop a percent frequency distribution.

2.17 Consider the frequency distribution of exam scores given below.

Class	Frequency
$90 < 100$	12
$80 < 90$	17
$70 < 80$	14
$60 < 70$	5
$50 < 60$	2

- **a** Develop a relative frequency distribution and a percent frequency distribution.
- **b** Develop a cumulative frequency distribution and a cumulative percent frequency distribution.
- **c** Draw a frequency polygon.
- **d** Draw a frequency ogive.

THE MARKETING RESEARCH CASE DS Design

Recall that 60 randomly selected shoppers have rated a new bottle design for a popular soft drink. The data are given below.

34	33	33	29	26	33	28	25	32	33
32	25	27	33	22	27	32	33	32	29
24	30	20	34	31	32	30	35	33	31
32	28	30	31	31	33	29	27	34	31
31	28	33	31	32	28	26	29	32	34
32	30	34	32	30	30	32	31	29	33

Use these data to work exercises 2.18 and 2.19.

2.18

- **a** Find the number of classes that should be used to construct a frequency distribution and histogram for the bottle design ratings.
- **b** If we round up to the nearest whole rating point, show that we should employ a class length equal to 3.
- **c** Define the nonoverlapping classes for a frequency distribution.
- **d** Tally the number of ratings in each class and develop a frequency distribution.
- **e** Draw the frequency histogram for the ratings data, and describe the distribution shape. DS Design

2.19

- **a** Construct a relative frequency distribution and a percent frequency distribution for the bottle design ratings.
- **b** Construct a cumulative frequency distribution and a cumulative percent frequency distribution.
- **c** Draw a frequency ogive for the bottle design ratings. DS Design

2.20 Table 2.11 gives the 25 most powerful celebrities and their annual pay as ranked by the editors of *Forbes* magazine and as listed on the Forbes.com website on June 14, 2011. DS PowerCeleb

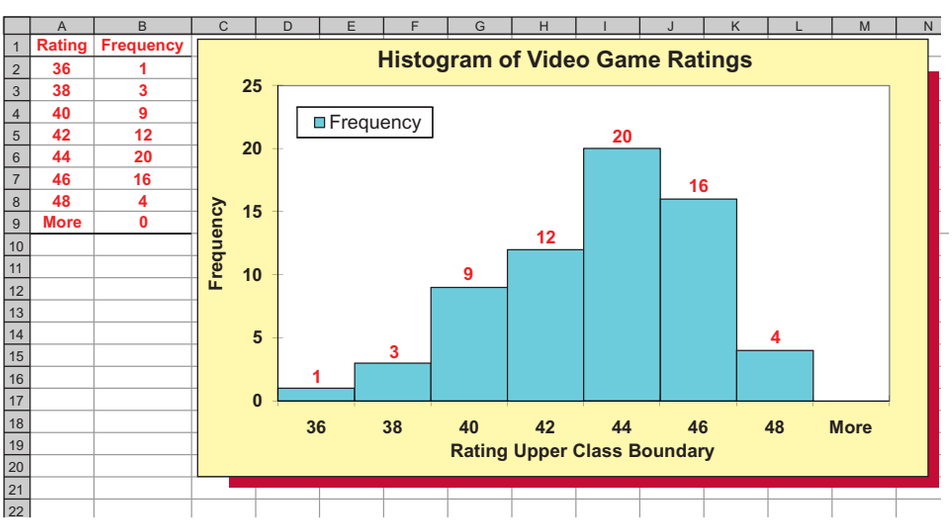
- **a** Develop a frequency distribution for the celebrity pay data and draw a histogram.
- **b** Develop a cumulative frequency distribution and a cumulative percent frequency distribution for the celebrity pay data.
- **c** Draw a percent frequency ogive for the celebrity pay data.

TABLE 2.11 The 25 Most Powerful Celebrities as Rated by *Forbes* Magazine DS PowerCeleb

Power Ranking	Celebrity Name	Pay ($mil)	Power Ranking	Celebrity Name	Pay ($mil)
1	Lady Gaga	90	14	Kobe Bryant	53
2	Oprah Winfrey	290	15	Leonardo Dicaprio	77
3	Justin Bieber	53	16	Black Eyed Peas	61
4	U2	195	17	Donald Trump	60
5	Elton John	100	18	Dr. Phil McGraw	80
6	Tiger Woods	75	19	Tyler Perry	130
7	Taylor Swift	45	20	Paul McCartney	67
8	Bon Jovi	125	21	Jennifer Aniston	28
9	Simon Cowell	90	22	Steven Spielberg	107
10	LeBron James	48	23	Rush Limbaugh	64
11	Angelina Jolie	30	24	Ryan Seacrest	61
12	Katy Perry	44	25	Roger Federer	47
13	Johnny Depp	50			

Source: http://www.forbes.com/wealth/celebrities (accessed June 14, 2011).

FIGURE 2.15 Excel Frequency Histogram of the 65 Satisfaction Ratings (for Exercise 2.21)



2.21 THE VIDEO GAME SATISFACTION RATING CASE DS VideoGame

Recall that Table 1.7 (page 13) presents the satisfaction ratings for the XYZ-Box video game system that have been given by 65 randomly selected purchasers. Figure 2.15 gives the Excel output of a histogram of these satisfaction ratings.

a Describe where the satisfaction ratings seem to be concentrated.
b Describe and interpret the shape of the distribution of ratings.
c Write out the eight classes used to construct this histogram.
d Construct a cumulative frequency distribution of the satisfaction ratings using the histogram classes.

2.22 THE BANK CUSTOMER WAITING TIME CASE DS WaitTime

Recall that Table 1.8 (page 13) presents the waiting times for teller service during peak business hours of 100 randomly selected bank customers. Figure 2.16 gives the MINITAB output of a histogram of these waiting times that has been constructed using automatic classes.

a Describe where the waiting times seem to be concentrated.
b Describe and interpret the shape of the distribution of waiting times.
c What is the class length that has been automatically defined by MINITAB?
d Write out the automatically defined classes and construct a cumulative percent frequency distribution of the waiting times using these classes.

FIGURE 2.16 MINITAB Frequency Histogram of the 100 Waiting Times Using Automatic Classes (for Exercise 2.22)

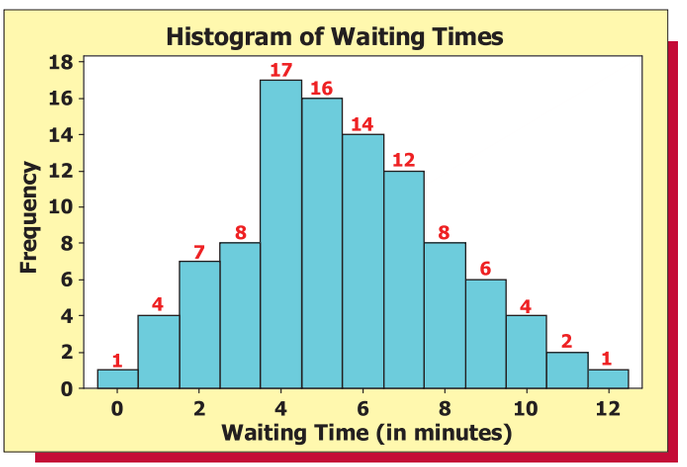
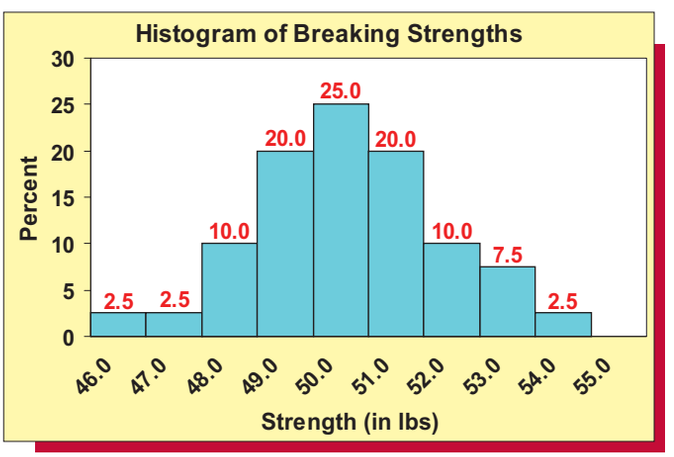


FIGURE 2.17 Percent Frequency Histogram of the 40 Breaking Strengths (for Exercise 2.23)



2.23 THE TRASH BAG CASE DS TrashBag

Recall that Table 1.9 (page 14) presents the breaking strengths of 40 trash bags selected during a 40-hour pilot production run. Figure 2.17 gives a percent frequency histogram of these breaking strengths.

a Describe where the breaking strengths seem to be concentrated.
b Describe and interpret the shape of the distribution of breaking strengths.
c What is the class length?
d Write out the classes and construct a percent frequency ogive for the breaking strengths using these classes.

2.24 Table 2.12 gives the franchise value and 2010 revenues for each of the 30 teams in Major League Baseball as reported by *Forbes* magazine and as listed on the Forbes.com website on June 14, 2011. DS MLBTeams

a Develop a frequency distribution and a frequency histogram for the 30 team values. Then describe the distribution of team values.
b Develop a percent frequency distribution and a percent frequency histogram for the 30 team revenues. Then describe the distribution of team revenues.
c Draw a percent frequency polygon for the 30 team values.

2.25 Table 2.13 gives America's top 40 best small companies of 2010 as rated on the Forbes.com website on June 14, 2011. DS SmallComp

a Develop a frequency distribution and a frequency histogram for the sales values. Describe the distribution of these sales values.
b Develop a percent frequency histogram for the sales growth values and then describe this distribution.

TABLE 2.12 Major League Baseball Team Valuations and Revenues as Given on the Forbes.com Website on June 14, 2011 (for Exercise 2.24) DS MLBTeams

Rank	Team	Value ($mil)	Revenue ($mil)	Rank	Team	Value ($mil)	Revenue ($mil)
1	New York Yankees	1,700	427	16	Washington Nationals	417	194
2	Boston Red Sox	912	272	17	Colorado Rockies	414	188
3	Los Angeles Dodgers	800	246	18	Baltimore Orioles	411	175
4	Chicago Cubs	773	258	19	San Diego Padres	406	159
5	New York Mets	747	233	20	Arizona Diamondbacks	396	180
6	Philadelphia Phillies	609	239	21	Detroit Tigers	385	192
7	San Francisco Giants	563	230	22	Milwaukee Brewers	376	179
8	Texas Rangers	561	206	23	Cincinnati Reds	375	179
9	Los Angeles Angels	554	222	24	Florida Marlins	360	143
10	Chicago White Sox	526	210	25	Cleveland Indians	353	168
11	St Louis Cardinals	518	207	26	Kansas City Royals	351	160
12	Minnesota Twins	490	213	27	Toronto Blue Jays	337	168
13	Atlanta Braves	482	201	28	Tampa Bay Rays	331	166
14	Houston Astros	474	197	29	Oakland Athletics	307	161
15	Seattle Mariners	449	204	30	Pittsburgh Pirates	304	160

Source: http://www.forbes.com/lists/2011/33/baseball-valuations-11_land.html (accessed June 14, 2011).

TABLE 2.13 America's Top 40 Best Small Companies of 2010 as Rated by *Forbes* Magazine (for Exercise 2.25) DS SmallComp

Rank	Company	Sales ($mil)	Sales Growth (%)	Rank	Company	Sales ($mil)	Sales Growth (%)
1	Medifast	218	41	21	TransDigm Group	802	22
2	InterDigital	359	17	22	Quality Systems	308	27
3	American Public Education	174	49	23	Continental Resources	906	15
4	Deckers Outdoor	869	33	24	Balchem	235	31
5	WebMD Health	480	28	25	HMS Holdings In	261	39
6	NutriSystem	534	63	26	Interactive Intelligence	143	21
7	National Presto Industries	491	27	27	VSE	960	42
8	Industrial Services of America	285	3	28	Under Armour	926	34
9	True Religion Apparel	335	54	29	F5 Networks	803	31
10	Transcend Services	84	33	30	InnerWorkings	439	64
11	Rackspace Hosting	698	51	31	Akamai Technologies	930	35
12	UFP Technologies	115	8	32	Lumber Liquidators	598	26
13	GeoResources	98	86	33	Allegiant Travel	604	47
14	Strayer Education	579	22	34	Steven Madden	570	7
15	Tempur Pedic International	986	4	35	KMG Chemicals	195	35
16	iRobot	373	27	36	Sapient	766	23
17	Dolby Laboratories	859	21	37	Concur Technologies	280	37
18	Hittite Microwave	200	24	38	Syntel	469	19
19	Capella Education	385	23	39	Atwood Oceanics	621	33
20	LoopNet	75	37	40	FactSet Research Systems	641	21

Source: http://www.forbes.com/lists/2010/23/best-small-companies-10_land.html (accessed June 14, 2011).

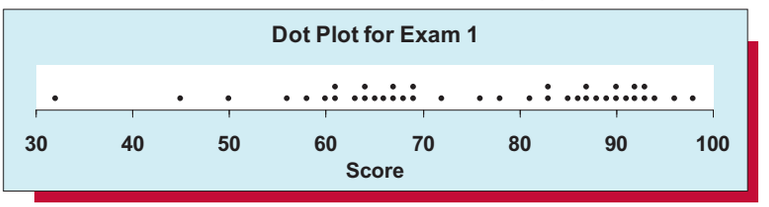
LO2-4 Construct and interpret dot plots.

2.3 Dot Plots

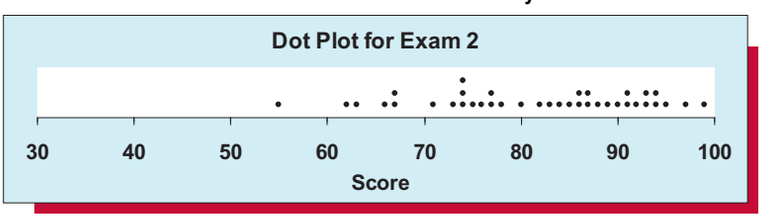
A very simple graph that can be used to summarize a data set is called a **dot plot.** To make a dot plot we draw a horizontal axis that spans the range of the measurements in the data set. We then place dots above the horizontal axis to represent the measurements. As an example, Figure 2.18(a) shows a dot plot of the exam scores in Table 2.8. Remember, these are the scores for the first exam given before implementing a strict attendance policy. The horizontal axis spans exam

FIGURE 2.18 Comparing Exam Scores Using Dot Plots

(a) Dot Plot of Scores on Exam 1: Before Attendance Policy



(b) Dot Plot of Scores on Exam 2: After Attendance Policy



scores from 30 to 100. Each dot above the axis represents an exam score. For instance, the two dots above the score of 90 tell us that two students received a 90 on the exam. The dot plot shows us that there are two concentrations of scores—those in the 80s and 90s and those in the 60s. Figure 2.18(b) gives a dot plot of the scores on the second exam (which was given after imposing the attendance policy). As did the percent frequency polygon for Exam 2 in Figure 2.13, this second dot plot shows that the attendance policy eliminated the concentration of scores in the 60s.

Dot plots are useful for detecting **outliers,** which are unusually large or small observations that are well separated from the remaining observations. For example, the dot plot for exam 1 indicates that the score 32 seems unusually low. How we handle an outlier depends on its cause. If the outlier results from a measurement error or an error in recording or processing the data, it should be corrected. If such an outlier cannot be corrected, it should be discarded. If an outlier is not the result of an error in measuring or recording the data, its cause may reveal important information. For example, the outlying exam score of 32 convinced the author that the student needed a tutor. After working with a tutor, the student showed considerable improvement on Exam 2. A more precise way to detect outliers is presented in Section 3.3.

Exercises for Section 2.3

CONCEPTS

connect

2.26 When we construct a dot plot, what does the horizontal axis represent? What does each dot represent?

2.27 If a data set consists of 1,000 measurements, would you summarize the data set using a histogram or a dot plot? Explain.

METHODS AND APPLICATIONS

2.28 The following data consist of the number of students who were absent in a professor's statistics class each day during the last month. DS AbsenceData

2	0	3	1	2	5	8	0	1	4
1	10	6	2	2	0	3	6	0	1

Construct a dot plot of these data, and then describe the distribution of absences.

2.29 The following are the revenue growth rates for 30 fast-growing companies. DS RevGrowth30

93%	43%	91%	49%	70%	44%	71%	70%	52%	59%
33%	40%	60%	35%	51%	48%	39%	61%	25%	87%
87%	46%	38%	30%	33%	43%	29%	38%	60%	32%

Develop a dot plot for these data and describe the distribution of revenue growth rates.

2.30 The yearly home run totals for Babe Ruth during his career as a New York Yankee are as follows (the totals are arranged in increasing order): 22, 25, 34, 35, 41, 41, 46, 46, 46, 47, 49, 54, 54, 59, 60. Construct a dot plot for these data and then describe the distribution of home run totals.
DS RuthsHomers

2.4 Stem-and-Leaf Displays

LO2-5 Construct and interpret stem-and-leaf displays.

Another simple graph that can be used to quickly summarize a data set is called a **stem-and-leaf display.** This kind of graph places the measurements in order from smallest to largest, and allows the analyst to simultaneously see all of the measurements in the data set and see the shape of the data set's distribution.

EXAMPLE 2.4 The Car Mileage Case: Estimating Mileage

Table 2.14 presents the sample of 50 gas mileages for the new midsize model previously introduced in Chapter 1. To develop a stem-and-leaf display, we note that the sample mileages range from 29.8 to 33.3 and we place the leading digits of these mileages—the whole numbers 29, 30, 31, 32, and 33—in a column on the left side of a vertical line as follows.

```
29 |
30 |
31 |
32 |
33 |
```

This vertical arrangement of leading digits forms the **stem** of the display. Next, we pass through the mileages in Table 2.14 one at a time and place each last digit (the tenths place) to the right of the vertical line in the row corresponding to its leading digits. For instance, the first three mileages—30.8, 31.7, and 30.1—are arranged as follows:

```
29 |
30 | 8 1
31 | 7
32 |
33 |
```

We form the **leaves** of the display by continuing this procedure as we pass through all 50 mileages. After recording the last digit for each of the mileages, we sort the digits in each row from smallest to largest and obtain the stem-and-leaf display that follows:

```
29 | 8
30 | 1 3 4 5 5 6 7 7 8 8 8
31 | 0 0 1 2 3 3 4 4 4 4 4 5 5 6 6 7 7 7 8 8 9 9
32 | 0 1 1 1 2 3 3 4 4 5 5 7 7 8
33 | 0 3
```

As we have said, the numbers to the left of the vertical line form the stem of the display. Each number to the right of the vertical line is a leaf. Each combination of a stem value and a leaf value

TABLE 2.14 A Sample of 50 Mileages for a New Midsize Model DS GasMiles

30.8	30.8	32.1	32.3	32.7
31.7	30.4	31.4	32.7	31.4
30.1	32.5	30.8	31.2	31.8
31.6	30.3	32.8	30.7	31.9
32.1	31.3	31.9	31.7	33.0
33.3	32.1	31.4	31.4	31.5
31.3	32.5	32.4	32.2	31.6
31.0	31.8	31.0	31.5	30.6
32.0	30.5	29.8	31.7	32.3
32.4	30.5	31.1	30.7	31.4

represents a measurement in the data set. For instance, the first row in the display

29 | 8

tells us that the first two digits are 29 and that the last (tenth place) digit is 8—that is, this combination represents the mileage 29.8 mpg. Similarly, the last row

33 | 0 3

represents the mileages 33.0 mpg and 33.3 mpg.

The entire stem-and-leaf display portrays the overall distribution of the sample mileages. It groups the mileages into classes, and it graphically illustrates how many mileages are in each class, as well as how the mileages are distributed within each class. The first class corresponds to the stem 29 and consists of the mileages from 29.0 to 29.9. There is one mileage—29.8—in this class. The second class corresponds to the stem 30 and consists of the mileages from 30.0 to 30.9. There are 11 mileages in this class. Similarly, the third, fourth, and fifth classes correspond to the stems 31, 32, and 33 and contain, respectively, 22 mileages, 14 mileages, and 2 mileages. Moreover, the stem-and-leaf display shows that the distribution of mileages is quite symmetrical. To see this, imagine turning the stem-and-leaf display on its side so that the vertical line becomes a horizontal number line. We see that the display now resembles a symmetrically shaped histogram. However, *the stem-and-leaf display is advantageous because it allows us to actually see the measurements in the data set* in addition to the distribution's shape.

When constructing a stem-and-leaf display, there are no rules that dictate the number of stem values (rows) that should be used. If we feel that the display has collapsed the mileages too closely together, we can stretch the display by assigning each set of leading digits to two or more rows. This is called *splitting the stems*. For example, in the following stem-and-leaf display of the mileages the first (uppermost) stem value of 30 is used to represent mileages between 30.0 and 30.4. The second stem value of 30 is used to represent mileages between 30.5 and 30.9.

```
29 | 8
30 | 1 3 4
30 | 5 5 6 7 7 8 8 8
31 | 0 0 1 2 3 3 4 4 4 4 4
31 | 5 5 6 6 7 7 7 8 8 9 9
32 | 0 1 1 1 2 3 3 4 4
32 | 5 5 7 7 8
33 | 0 3
```

Notice that, in this particular case, splitting the stems produces a display that seems to more clearly reveal the symmetrical shape of the distribution of mileages.

Most statistical software packages can be used to construct stem-and-leaf displays. Figure 2.19 gives a MINITAB stem-and-leaf display of the 50 sample mileages. This output has been obtained by splitting the stems—MINITAB produced this display automatically. MINITAB also provides

FIGURE 2.19 MINITAB Stem-and-Leaf Display of the 50 Mileages

```
Stem-and-Leaf Display: Mpg
Stem-and-leaf of Mpg N = 50
Leaf unit = 0.10
 1    29  8
 4    30  134
 12   30  55677888
 23   31  00123344444
(11)  31  55667778899
 16   32  011123344
 7    32  55778
 2    33  03
```

an additional column of numbers (on the left) that provides information about how many mileages are in the various rows. For example, if we look at the MINITAB output, the 11 (in parentheses) tells us that there are 11 mileages between 31.5 mpg and 31.9 mpg. The 12 (no parentheses) tells us that a total of 12 mileages are at or below 30.9 mpg, while the 7 tells us that a total of 7 mileages are at or above 32.5 mpg.

It is possible to construct a stem-and-leaf display from measurements containing any number of digits. To see how this can be done, consider the following data which consists of the number of DVD players sold by an electronics manufacturer for each of the last 12 months.

13,502	15,932	14,739	15,249	14,312	17,111	DS DVDPlayers
19,010	16,121	16,708	17,886	15,665	16,475	

To construct a stem-and-leaf display, we will use only the first three digits of each sales value and we will define leaf values consisting of one digit. The stem will consist of the values 13, 14, 15, 16, 17, 18, and 19 (which represent thousands of units sold). Each leaf will represent the remaining three digits rounded to the nearest 100 units sold. For example, 13,502 will be represented by placing the leaf value 5 in the row corresponding to 13. To express the fact that the leaf 5 represents 500, we say that the **leaf unit** is 100. Using this procedure, we obtain the following stem-and-leaf display:

Leaf unit = 100

13	5
14	3 7
15	2 7 9
16	1 5 7
17	1 9
18	
19	0

The standard practice of always using a single digit for each leaf allows us to construct a stem-and-leaf display for measurements having any number of digits as long as we appropriately define a leaf unit. However, it is not possible to recover the original measurements from such a display. If we do not have the original measurements, the best we can do is to approximate them by multiplying the digits in the display by the leaf unit. For instance, the measurements in the row corresponding to the stem value 17 can be approximated to be $171 \times (100) = 17{,}100$ and $179 \times (100) = 17{,}900$. In general, leaf units can be any power of 10 such as 0.1, 1, 10, 100, 1000, and so on. If no leaf unit is given for a stem-and-leaf display, we assume its value is 1.0.

We summarize how to set up a stem-and-leaf display in the following box:

Constructing a Stem-and-Leaf Display

1. Decide what units will be used for the stems and the leaves. Each leaf must be a single digit and the stem values will consist of appropriate leading digits. As a general rule, there should be between 5 and 20 stem values.
2. Place the stem values in a column to the left of a vertical line with the smallest value at the top of the column and the largest value at the bottom.
3. To the right of the vertical line, enter the leaf for each measurement into the row corresponding to the proper stem value. Each leaf should be a single digit—these can be rounded values that were originally more than one digit if we are using an appropriately defined leaf unit.
4. Rearrange the leaves so that they are in increasing order from left to right.

If we wish to compare two distributions, it is convenient to construct a **back-to-back stem-and-leaf display.** Figure 2.20 presents a back-to-back stem-and-leaf display for the previously discussed exam scores. The left side of the display summarizes the scores for the first exam. Remember, this exam was given before implementing a strict attendance policy. The right side of the display summarizes the scores for the second exam (which was given after imposing the attendance policy). Looking at the left side of the display, we see that for the first exam there are two concentrations of scores—those in the 80s and 90s and those in the 60s. The right side of the

FIGURE 2.20 A Back-to-Back Stem-and-Leaf Display of the Exam Scores

Exam 1		Exam 2
2	3	
	3	
	4	
5	4	
0	5	
8 6	5	5
4 4 3 1 1 0	6	2 3
9 9 8 7 7 6 5	6	6 7 7
2	7	1 3 4 4 4
8 6	7	5 6 7 7 8
3 3 1	8	0 2 3 4
9 8 7 7 6 5	8	5 6 6 7 7 8 9
4 3 3 2 2 1 0 0	9	0 1 1 2 3 3 4 4
8 6	9	5 7 9

display shows that the attendance policy eliminated the concentration of scores in the 60s and illustrates that the scores on exam 2 are almost single peaked and somewhat skewed to the left.

Stem-and-leaf displays are useful for detecting **outliers,** which are unusually large or small observations that are well separated from the remaining observations. For example, the stem-and-leaf display for exam 1 indicates that the score 32 seems unusually low. How we handle an outlier depends on its cause. If the outlier results from a measurement error or an error in recording or processing the data, it should be corrected. If such an outlier cannot be corrected, it should be discarded. If an outlier is not the result of an error in measuring or recording the data, its cause may reveal important information. For example, the outlying exam score of 32 convinced the author that the student needed a tutor. After working with a tutor, the student showed considerable improvement on Exam 2. A more precise way to detect outliers is presented in Section 3.3.

Exercises for Section 2.4

CONCEPTS

2.31 Explain the difference between a histogram and a stem-and-leaf display.

2.32 What are the advantages of using a stem-and-leaf display?

2.33 If a data set consists of 1,000 measurements, would you summarize the data set by using a stem-and-leaf display or a histogram? Explain.

METHODS AND APPLICATIONS

2.34 The following data consist of the revenue growth rates (in percent) for a group of 20 firms. Construct a stem-and-leaf display for these data. DS RevGrowth20

36	59	42	65	91	32	56	28	49	51
30	55	33	63	70	44	42	83	53	43

2.35 The following data consist of the profit margins (in percent) for a group of 20 firms. Construct a stem-and-leaf display for these data. DS ProfitMar20

25.2	16.1	22.2	15.2	14.1	15.2	14.4	15.9	10.4	14.0
16.4	13.9	10.4	13.8	14.9	16.1	15.8	13.2	16.8	12.6

2.36 The following data consist of the sales figures (in millions of dollars) for a group of 20 firms. Construct a stem-and-leaf display for these data. Use a leaf unit equal to 100. DS Sales20

6835	1973	2820	5358	1233	3291	2707	3291	2675	3707
3517	1449	2384	1376	1725	6047	7903	4616	1541	4189

2.37 THE e-BILLING CASE AND THE MARKETING RESEARCH CASE

Figure 2.21 gives stem-and-leaf displays of the payment times in Table 2.4 on page 42 and of the bottle design ratings in Table 1.5 on page 10. Describe the shapes of the two displays. DS Paytime DS Design

FIGURE 2.21 Stem-and-Leaf Displays of the Payment Times and Bottle Design Ratings (for Exercise 2.37)

Payment Times		Ratings	
10	0	20	0
11		21	
12	00	22	0
13	000	23	
14	0000	24	0
15	0000000	25	00
16	000000000	26	00
17	00000000	27	000
18	000000	28	0000
19	00000	29	00000
20	000	30	000000
21	0000	31	00000000
22	000	32	00000000000
23	00	33	0000000000
24	000	34	00000
25	00	35	0
26	0		
27	0		
28			
29	0		

FIGURE 2.22 Stem-and-Leaf Display of the 40 Breaking Strengths (for Exercise 2.38)

Stem-and-leaf plot for strength
Stem unit = 1 Leaf unit = 0.1

Frequency	Stem	Leaf
1	46	8
0	47	
1	47	5
2	48	2 3
2	48	5 8
4	49	0 2 3 4
4	49	5 6 8 9
4	50	0 1 2 3
6	50	5 6 7 8 9 9
5	51	0 1 2 3 4
3	51	5 7 9
2	52	0 3
2	52	5 6
2	53	0 2
1	53	5
1	54	0
40		

2.38 THE TRASH BAG CASE DS TrashBag

Figure 2.22 gives a stem-and-leaf display of the sample of 40 breaking strengths in the trash bag case.

a Use the stem-and-leaf display to describe the distribution of breaking strengths.

b Write out the 10 smallest breaking strengths as they would be expressed in the original data.

2.39 Babe Ruth's record of 60 home runs in a single year was broken by Roger Maris, who hit 61 home runs in 1961. The yearly home run totals for Ruth in his career as a New York Yankee are (arranged in increasing order) 22, 25, 34, 35, 41, 41, 46, 46, 46, 47, 49, 54, 54, 59, and 60. The yearly home run totals for Maris over his career in the American League are (arranged in increasing order) 8, 13, 14, 16, 23, 26, 28, 33, 39, and 61. Compare Ruth's and Maris's home run totals by constructing a back-to-back stem-and-leaf display. What would you conclude about Maris's record-breaking year? DS HomeRuns

2.40 THE BANK CUSTOMER WAITING TIME CASE DS WaitTime

Table 2.15 reproduces the 100 waiting times for teller service that were originally given in Table 1.8 (page 13).

a Construct a stem-and-leaf display of the waiting times.

b Describe the distribution of the waiting times.

2.41 THE VIDEO GAME SATISFACTION RATING CASE DS VideoGame

Recall that 65 purchasers have participated in a survey and have rated the XYZ-Box video game system. The composite ratings that have been obtained are as follows:

39	38	40	40	40	46	43	38	44	44	44
45	42	42	47	46	45	41	43	46	44	42
38	46	45	44	41	45	40	36	48	44	47
42	44	44	43	43	46	43	44	44	46	43
42	40	42	45	39	43	44	44	41	39	45
41	39	46	45	43	47	41	45	45	41	

a Construct a stem-and-leaf display for the 65 composite ratings. Hint: Each whole number rating can be written with an "implied tenth place" of zero. For instance, 39 can be written as 39.0. Use the implied zeros as the leaf values and the whole numbers 36, 37, 38, 39, etc. as the stem values.

b Describe the distribution of composite ratings.

c If we consider a purchaser to be "very satisfied" if his or her composite score is at least 42, can we say that almost all purchasers of the XYZ-Box video game system are "very satisfied"?

TABLE 2.15 Waiting Times (in Minutes) for the Bank Customer Waiting Time Case (for Exercise 2.40) DS WaitTime

1.6	6.2	3.2	5.6	7.9	6.1	7.2
6.6	5.4	6.5	4.4	1.1	3.8	7.3
5.6	4.9	2.3	4.5	7.2	10.7	4.1
5.1	5.4	8.7	6.7	2.9	7.5	6.7
3.9	.8	4.7	8.1	9.1	7.0	3.5
4.6	2.5	3.6	4.3	7.7	5.3	6.3
6.5	8.3	2.7	2.2	4.0	4.5	4.3
6.4	6.1	3.7	5.8	1.4	4.5	3.8
8.6	6.3	.4	8.6	7.8	1.8	5.1
4.2	6.8	10.2	2.0	5.2	3.7	5.5
5.8	9.8	2.8	8.0	8.4	4.0	
3.4	2.9	11.6	9.5	6.3	5.7	
9.3	10.9	4.3	1.3	4.4	2.4	
7.4	4.7	3.1	4.8	5.2	9.2	
1.8	3.9	5.8	9.9	7.4	5.0	

2.5 Contingency Tables (Optional)

LO2-6 Examine the relationships between variables by using contingency tables (Optional).

Previous sections in this chapter have presented methods for summarizing data for a single variable. Often, however, we wish to use statistics to study possible relationships between several variables. In this section we present a simple way to study the relationship between two variables. Crosstabulation is a process that classifies data on two dimensions. This process results in a table that is called a **contingency table**. Such a table consists of rows and columns—the rows classify the data according to one dimension and the columns classify the data according to a second dimension. Together, the rows and columns represent all possibilities (or *contingencies*).

EXAMPLE 2.5 The Brokerage Firm Case: Studying Client Satisfaction

An investment broker sells several kinds of investment products—a stock fund, a bond fund, and a tax-deferred annuity. The broker wishes to study whether client satisfaction with its products and services depends on the type of investment product purchased. To do this, 100 of the broker's clients are randomly selected from the population of clients who have purchased shares in exactly one of the funds. The broker records the fund type purchased by each client and has one of its investment counselors personally contact the client. When contacted, the client is asked to rate his or her level of satisfaction with the purchased fund as high, medium, or low. The resulting data are given in Table 2.16.

Looking at the raw data in Table 2.16, it is difficult to see whether the level of client satisfaction varies depending on the fund type. We can look at the data in an organized way by constructing a contingency table. A cross-tabulation of fund type versus level of client satisfaction is shown in Table 2.17. The classification categories for the two variables are defined along the left and top margins of the table. The three row labels—bond fund, stock fund, and tax deferred annuity—define the three fund categories and are given in the left table margin. The three column labels—high, medium, and low—define the three levels of client satisfaction and are given along the top table margin. Each row and column combination, that is, each fund type and level of satisfaction combination, defines what we call a "cell" in the table. Because each of the randomly selected clients has invested in exactly one fund type and has reported exactly one level of satisfaction, each client can be placed in a particular cell in the contingency table. For example, because client number 1 in Table 2.16 has invested in the bond fund and reports a high level of client satisfaction, client number 1 can be placed in the upper left cell of the table (the cell defined by the Bond Fund row and High Satisfaction column).

We fill in the cells in the table by moving through the 100 randomly selected clients and by tabulating the number of clients who can be placed in each cell. For instance, moving through the 100 clients results in placing 15 clients in the "bond fund—high" cell, 12 clients in the

TABLE 2.16 Results of a Customer Satisfaction Survey Given to 100 Randomly Selected Clients Who Invest in One of Three Fund Types—a Bond Fund, a Stock Fund, or a Tax-Deferred Annuity DS Invest

Client	Fund Type	Level of Satisfaction	Client	Fund Type	Level of Satisfaction	Client	Fund Type	Level of Satisfaction
1	BOND	HIGH	35	STOCK	HIGH	69	BOND	MED
2	STOCK	HIGH	36	BOND	MED	70	TAXDEF	MED
3	TAXDEF	MED	37	TAXDEF	MED	71	TAXDEF	MED
4	TAXDEF	MED	38	TAXDEF	LOW	72	BOND	HIGH
5	STOCK	LOW	39	STOCK	HIGH	73	TAXDEF	MED
6	STOCK	HIGH	40	TAXDEF	MED	74	TAXDEF	LOW
7	STOCK	HIGH	41	BOND	HIGH	75	STOCK	HIGH
8	BOND	MED	42	BOND	HIGH	76	BOND	HIGH
9	TAXDEF	LOW	43	BOND	LOW	77	TAXDEF	LOW
10	TAXDEF	LOW	44	TAXDEF	LOW	78	BOND	MED
11	STOCK	MED	45	STOCK	HIGH	79	STOCK	HIGH
12	BOND	LOW	46	BOND	HIGH	80	STOCK	HIGH
13	STOCK	HIGH	47	BOND	MED	81	BOND	MED
14	TAXDEF	MED	48	STOCK	HIGH	82	TAXDEF	MED
15	TAXDEF	MED	49	TAXDEF	MED	83	BOND	HIGH
16	TAXDEF	LOW	50	TAXDEF	MED	84	STOCK	MED
17	STOCK	HIGH	51	STOCK	HIGH	85	STOCK	HIGH
18	BOND	HIGH	52	TAXDEF	MED	86	BOND	MED
19	BOND	MED	53	STOCK	HIGH	87	TAXDEF	MED
20	TAXDEF	MED	54	TAXDEF	MED	88	TAXDEF	LOW
21	TAXDEF	MED	55	STOCK	LOW	89	STOCK	HIGH
22	BOND	HIGH	56	BOND	HIGH	90	TAXDEF	MED
23	TAXDEF	MED	57	STOCK	HIGH	91	BOND	HIGH
24	TAXDEF	LOW	58	BOND	MED	92	TAXDEF	HIGH
25	STOCK	HIGH	59	TAXDEF	LOW	93	TAXDEF	LOW
26	BOND	HIGH	60	TAXDEF	LOW	94	TAXDEF	LOW
27	TAXDEF	LOW	61	STOCK	MED	95	STOCK	HIGH
28	BOND	MED	62	BOND	LOW	96	BOND	HIGH
29	STOCK	HIGH	63	STOCK	HIGH	97	BOND	MED
30	STOCK	HIGH	64	TAXDEF	MED	98	STOCK	HIGH
31	BOND	MED	65	TAXDEF	MED	99	TAXDEF	MED
32	TAXDEF	MED	66	TAXDEF	LOW	100	TAXDEF	MED
33	BOND	HIGH	67	STOCK	HIGH			
34	STOCK	MED	68	BOND	HIGH			

TABLE 2.17 A Contingency Table of Fund Type versus Level of Client Satisfaction

	Level of Satisfaction			
Fund Type	**High**	**Medium**	**Low**	**Total**
Bond Fund	15	12	3	30
Stock Fund	24	4	2	30
Tax Deferred Annuity	1	24	15	40
Total	40	40	20	100

"bond fund—medium" cell, and so forth. The counts in the cells are called the **cell frequencies.** In Table 2.17 these frequencies tell us that 15 clients invested in the bond fund and reported a high level of satisfaction, 4 clients invested in the stock fund and reported a medium level of satisfaction, and so forth.

The far right column in the table (labeled Total) is obtained by summing the cell frequencies across the rows. For instance, these totals tell us that $15 + 12 + 3 = 30$ clients invested in the bond fund, $24 + 4 + 2 = 30$ clients invested in the stock fund, and $1 + 24 + 15 = 40$ clients invested in the tax deferred annuity. These **row totals** provide a frequency distribution for the

different fund types. By dividing the row totals by the total of 100 clients surveyed, we can obtain relative frequencies; and by multiplying each relative frequency by 100, we can obtain percent frequencies. That is, we can obtain the frequency, relative frequency, and percent frequency distributions for fund type as follows:

Fund Type	Frequency	Relative Frequency	Percent Frequency
Bond fund	30	30/100 = .30	.30 (100) = 30%
Stock fund	30	30/100 = .30	.30 (100) = 30%
Tax deferred annuity	40	40/100 = .40	.40 (100) = 40%
	100		

We see that 30 percent of the clients invested in the bond fund, 30 percent invested in the stock fund, and 40 percent invested in the tax deferred annuity.

The bottom row in the table (labeled Total) is obtained by summing the cell frequencies down the columns. For instance, these totals tell us that $15 + 24 + 1 = 40$ clients reported a high level of satisfaction, $12 + 4 + 24 = 40$ clients reported a medium level of satisfaction, and $3 + 2 + 15 = 20$ clients reported a low level of satisfaction. These **column totals** provide a frequency distribution for the different satisfaction levels (see below). By dividing the column totals by the total of 100 clients surveyed, we can obtain relative frequencies, and by multiplying each relative frequency by 100, we can obtain percent frequencies. That is, we can obtain the frequency, relative frequency, and percent frequency distributions for level of satisfaction as follows:

Level of Satisfaction	Frequency	Relative Frequency	Percent Frequency
High	40	40/100 = .40	.40 (100) = 40%
Medium	40	40/100 = .40	.40 (100) = 40%
Low	20	20/100 = .20	.20 (100) = 20%
	100		

We see that 40 percent of all clients reported high satisfaction, 40 percent reported medium satisfaction, and 20 percent reported low satisfaction.

We have seen that the totals in the margins of the contingency table give us frequency distributions that provide information about each of the variables *fund type* and *level of client satisfaction*. However, the main purpose of constructing the table is to investigate possible relationships *between* these variables. Looking at Table 2.17, we see that clients who have invested in the stock fund seem to be highly satisfied and that those who have invested in the bond fund seem to have a high to medium level of satisfaction. However, clients who have invested in the tax deferred annuity seem to be less satisfied.

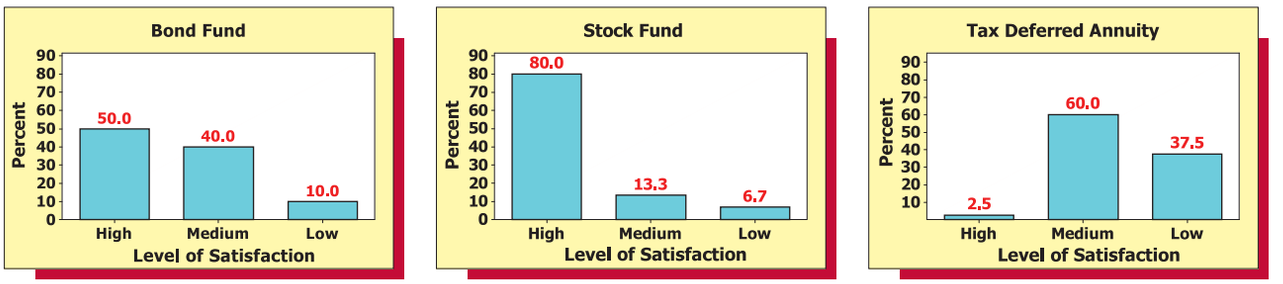
One good way to investigate relationships such as these is to compute **row percentages** and **column percentages.** We compute row percentages by dividing each cell's frequency by its corresponding row total and by expressing the resulting fraction as a percentage. For instance, the row percentage for the upper left-hand cell (bond fund and high level of satisfaction) in Table 2.17 is $(15/30) \times 100\% = 50\%$. Similarly, column percentages are computed by dividing each cell's frequency by its corresponding column total and by expressing the resulting fraction as a percentage. For example, the column percentage for the upper left-hand cell in Table 2.17 is $(15/40) \times 100\% = 37.5\%$. Table 2.18 summarizes all of the row percentages for the different fund types in Table 2.17. We see that each row in Table 2.18 gives a percentage frequency distribution of level of client satisfaction given a particular fund type.

For example, the first row in Table 2.18 gives a percent frequency distribution of client satisfaction for investors who have purchased shares in the bond fund. We see that 50 percent of

TABLE 2.18 Row Percentages for Each Fund Type

	Level of Satisfaction			
Fund Type	**High**	**Medium**	**Low**	**Total**
Bond Fund	50%	40%	10%	100%
Stock Fund	80%	13.33%	6.67%	100%
Tax Deferred	2.5%	60%	37.5%	100%

FIGURE 2.23 Bar Charts Illustrating Percent Frequency Distributions of Client Satisfaction as Given by the Row Percentages for the Three Fund Types in Table 2.18



bond fund investors report high satisfaction, while 40 percent of these investors report medium satisfaction, and only 10 percent report low satisfaction. The other rows in Table 2.18 provide percent frequency distributions of client satisfaction for stock fund and annuity purchasers.

All three percent frequency distributions of client satisfaction—for the bond fund, the stock fund, and the tax deferred annuity—are illustrated using bar charts in Figure 2.23. In this figure, the bar heights for each chart are the respective row percentages in Table 2.18. For example, these distributions tell us that 80 percent of stock fund investors report high satisfaction, while 97.5 percent of tax deferred annuity purchasers report medium or low satisfaction. Looking at the entire table of row percentages (or the bar charts in Figure 2.23), we might conclude that stock fund investors are highly satisfied, that bond fund investors are quite satisfied (but, somewhat less so than stock fund investors), and that tax-deferred-annuity purchasers are less satisfied than either stock fund or bond fund investors. In general, row percentages and column percentages help us to quantify relationships such as these.

In the investment example, we have cross-tabulated two qualitative variables. We can also cross-tabulate a quantitative variable versus a qualitative variable or two quantitative variables against each other. If we are cross-tabulating a quantitative variable, we often define categories by using appropriate ranges. For example, if we wished to cross-tabulate level of education (grade school, high school, college, graduate school) versus income, we might define income classes \$0–\$50,000, \$50,001–\$100,000, \$100,001–\$150,000, and above \$150,000.

Exercises for Section 2.5

CONCEPTS

connect™

2.42 Explain the purpose behind constructing a contingency table.

2.43 A contingency table consists of several "cells". Explain how we fill the cells in the table.

2.44 Explain how to compute (1) the row percentages for a contingency table, and (2) the column percentages. What information is provided by the row percentages in a particular row of the table? What information is provided by the column percentages in a particular column of the table?

METHODS AND APPLICATIONS

Exercises 2.45 through 2.47 are based on the following situation:

The marketing department at the Rola-Cola Bottling Company is investigating the attitudes and preferences of consumers toward Rola-Cola and a competing soft drink, Koka-Cola. Forty randomly selected shoppers are given a "blind taste-test" and are asked to give their cola preferences. The results are given in Table 2.19—each shopper's preference, Rola-Cola or Koka-Cola, is revealed to the shopper only after he or she has tasted both brands without knowing which cola is which. In addition, each survey participant is asked to answer three more questions: (1) Have you previously purchased Rola-Cola: Yes

TABLE 2.19 Rola-Cola Bottling Company Survey Results ColaSurvey

Shopper	Cola Preference	Previously Purchased?	Sweetness Preference	Monthly Cola Consumption	Shopper	Cola Preference	Previously Purchased?	Sweetness Preference	Monthly Cola Consumption
1	Koka	No	Very Sweet	4	21	Koka	No	Very Sweet	4
2	Rola	Yes	Sweet	8	22	Rola	Yes	Not So Sweet	9
3	Koka	No	Not So Sweet	2	23	Rola	Yes	Not So Sweet	3
4	Rola	Yes	Sweet	10	24	Koka	No	Not So Sweet	2
5	Rola	No	Very Sweet	7	25	Koka	No	Sweet	5
6	Rola	Yes	Not So Sweet	6	26	Rola	Yes	Very Sweet	7
7	Koka	No	Very Sweet	4	27	Koka	No	Very Sweet	7
8	Rola	No	Very Sweet	3	28	Rola	Yes	Sweet	8
9	Koka	No	Sweet	3	29	Rola	Yes	Not So Sweet	6
10	Rola	No	Very Sweet	5	30	Koka	No	Not So Sweet	3
11	Rola	Yes	Sweet	7	31	Koka	Yes	Sweet	10
12	Rola	Yes	Not So Sweet	13	32	Rola	Yes	Very Sweet	8
13	Rola	Yes	Very Sweet	6	33	Koka	Yes	Sweet	4
14	Koka	No	Very Sweet	2	34	Rola	No	Sweet	5
15	Koka	No	Not So Sweet	7	35	Rola	Yes	Not So Sweet	3
16	Rola	Yes	Sweet	9	36	Koka	No	Very Sweet	11
17	Koka	No	Not So Sweet	1	37	Rola	Yes	Not So Sweet	9
18	Rola	Yes	Very Sweet	5	38	Rola	No	Very Sweet	6
19	Rola	No	Sweet	4	39	Koka	No	Not So Sweet	2
20	Rola	No	Sweet	12	40	Rola	Yes	Sweet	5

or No? (2) What is your sweetness preference for cola drinks: very sweet, sweet, or not so sweet? (3) How many 12-packs of cola drinks does your family consume in a typical month? These responses are also given in Table 2.19.

2.45 Construct a contingency table using cola preference (Rola or Koka) as the row variable and Rola-Cola purchase history (Yes or No) as the column variable. Based on the table, answer the following.

- **a** How many shoppers who preferred Rola-Cola in the blind taste test had previously purchased Rola-Cola?
- **b** How many shoppers who preferred Koka-Cola in the blind taste test had not previously purchased Rola-Cola?
- **c** What kind of relationship, if any, seems to exist between cola preference and Rola-Cola purchase history?

2.46 Construct a contingency table using cola preference (Rola or Koka) as the row variable and sweetness preference (very sweet, sweet, or not so sweet) as the column variable. Based on the table, answer the following:

- **a** How many shoppers who preferred Rola-Cola in the blind taste test said that they preferred a cola drink to be either very sweet or sweet?
- **b** How many shoppers who preferred Koka-Cola in the blind taste test said that they preferred a cola drink to be not so sweet?
- **c** What kind of relationship, if any, seems to exist between cola preference and sweetness preference?

2.47 Construct a contingency table using cola preference (Rola or Koka) as the row variable and the number of 12-packs consumed in a typical month (categories 0 through 5, 6 through 10, and more than 10) as the column variable. Based on the table, answer the following:

- **a** How many shoppers who preferred Rola-Cola in the blind taste test purchase 10 or fewer 12-packs of cola drinks in a typical month?
- **b** How many shoppers who preferred Koka-Cola in the blind taste test purchase 6 or more 12-packs of cola drinks in a typical month?
- **c** What kind of relationship, if any, seems to exist between cola preference and cola consumption in a typical month?

2.48 A marketing research firm wishes to study the relationship between wine consumption and whether a person likes to watch professional tennis on television. One hundred randomly selected

people are asked whether they drink wine and whether they watch tennis. The following results are obtained: DS WineCons

	Watch Tennis	Do Not Watch Tennis	Total
Drink Wine	16	24	40
Do Not Drink Wine	4	56	60
Total	20	80	100

a What percentage of those surveyed both watch tennis and drink wine? What percentage of those surveyed do neither?
b Using the survey data, construct a table of row percentages.
c Using the survey data, construct a table of column percentages.
d What kind of relationship, if any, seems to exist between whether or not a person watches tennis and whether or not a person drinks wine?
e Illustrate your conclusion of part (d) by plotting bar charts of appropriate column percentages for people who watch tennis and for people who do not watch tennis.

2.49 In a survey of 1,000 randomly selected U.S. citizens aged 21 years or older, 721 believed that the amount of violent television programming had increased over the past 10 years, 454 believed that the overall quality of television programming had gotten worse over the past 10 years, and 362 believed both.

a Use this information to fill in the contingency table below.

	TV Violence Increased	TV Violence Not Increased	Total
TV Quality Worse			
TV Quality Not Worse			
Total			

b Using the completed contingency table, construct a table of row percentages.
c Using the completed contingency table, construct a table of column percentages.
d What kind of relationship, if any, seems to exist between whether a person believed that TV violence had increased over the past ten years and whether a person believed that the overall quality of TV programming had gotten worse over the past ten years?
e Illustrate your answer to part (d) by constructing bar charts of appropriate row percentages.

In Exercises 2.50 and 2.51 we consider the results of a Gallup Lifestyle Poll about restaurant tipping habits as reported by the Gallup News Service on January 8, 2007. The poll asked Americans to recommend the percentage of a restaurant bill that should be left as a tip. As reported on galluppoll.com, Americans gave an overall (average) recommendation of 16.2 percent.

2.50 As part of its study, Gallup investigated a possible relationship between tipping attitudes and income. Using the poll results, the following row percentages can be obtained for three income ranges—less than $30,000; $30,000 through $74,999; and $75,000 or more. DS RowPercents

	Appropriate Tip Percent*				
Income	**Less than 15%**	**15%**	**16–19%**	**20% or more**	**Total**
Less than $30,000	28.41%	42.04%	1.14%	28.41%	100%
$30,000 through $74,999	15.31%	42.86%	6.12%	35.71%	100%
$75,000 or more	8.16%	32.66%	9.18%	50.00%	100%

*Among those surveyed having an opinion.

a Construct a percentage bar chart of recommended tip percentage for each of the three income ranges.
b Using the bar charts, describe the relationship between recommended tip percentage and income level.

2.51 Gallup also asked survey participants if they have ever eaten at a restaurant and left no tip at all because of poor service. Almost half (46 percent) of those surveyed said they have done so. Because it seems that whether or not a person has left a restaurant without tipping might be related to a person's generosity in terms of recommended tip percentage, the survey investigated this possible relationship. Using the poll results, the column percentages at the top of page 67 can be obtained for each of the categories "Yes, have left without tipping" and "No, have not left without tipping." DS ColPercents

	Tip less than 15%	Tip 15% through 19%	Tip 20% or more
Yes, have left without tipping	64%	50%	35%
No, have not left without tipping	36%	50%	65%
Total	100%	100%	100%

a Construct a percentage bar chart of the categories "Yes, have left without tipping" and "No, have not left without tipping" for each of the tip categories "less than 15%," "15% through 19%," and "20% or more."

b Using the bar charts, describe the relationship between whether or not a person has left without tipping and tipping generosity.

2.6 Scatter Plots (Optional)

LO2-7 Examine the relationships between variables by using scatter plots (Optional).

We often study relationships between variables by using graphical methods. A simple graph that can be used to study the relationship between two variables is called a **scatter plot.** As an example, suppose that a marketing manager wishes to investigate the relationship between the sales volume (in thousands of units) of a product and the amount spent (in units of $10,000) on advertising the product. To do this, the marketing manager randomly selects 10 sales regions having equal sales potential. The manager assigns a different level of advertising expenditure for January 2011 to each sales region as shown in Table 2.20. At the end of the month, the sales volume for each region is recorded as also shown in Table 2.20.

A scatter plot of these data is given in Figure 2.24. To construct this plot, we place the variable advertising expenditure (denoted x) on the horizontal axis and we place the variable sales volume (denoted y) on the vertical axis. For the first sales region, advertising expenditure equals 5 and sales volume equals 89. We plot the point with coordinates $x = 5$ and $y = 89$ on the scatter plot to represent this sales region. Points for the other sales regions are plotted similarly. The scatter plot shows that there is a positive relationship between advertising expenditure and sales volume—that is, higher values of sales volume are associated with higher levels of advertising expenditure.

We have drawn a straight line through the plotted points of the scatter plot to represent the relationship between advertising expenditure and sales volume. We often do this when the relationship between two variables appears to be a **straight line,** or **linear.** Of course, the relationship between x and y in Figure 2.24 is not perfectly linear—not all of the points in the scatter plot are exactly on the line. Nevertheless, because the relationship between x and y appears to be approximately linear, it seems reasonable to represent the general relationship between these variables using a straight line. In future chapters we will explain ways to quantify such a relationship—that is, describe such a relationship numerically. We will show that we can statistically express the strength of a linear relationship and that we can calculate the equation of the line that best fits the points of a scatter plot.

TABLE 2.20 Values of Advertising Expenditure (in $10,000s) and Sales Volume (in 1000s) for Ten Sales Regions DS SalesPlot

Sales Region	Advertising Expenditure, x	Sales Volume, y
1	5	89
2	6	87
3	7	98
4	8	110
5	9	103
6	10	114
7	11	116
8	12	110
9	13	126
10	14	130

FIGURE 2.24 A Scatter Plot of Sales Volume versus Advertising Expenditure

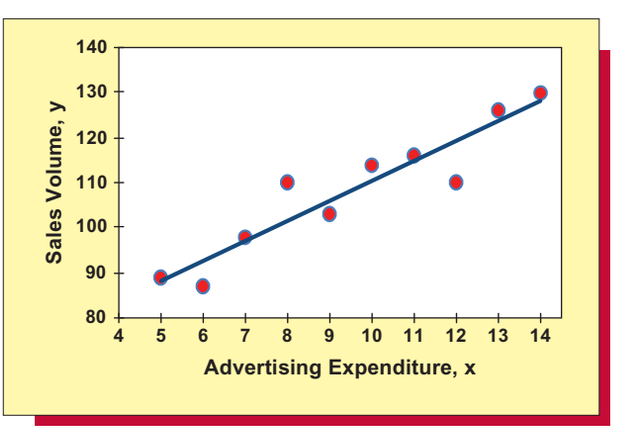


FIGURE 2.25 A Positive Linear Relationship

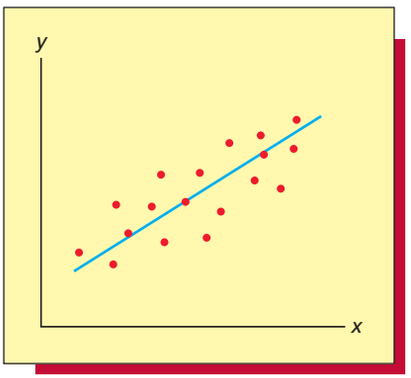


FIGURE 2.26 Little or No Linear Relationship

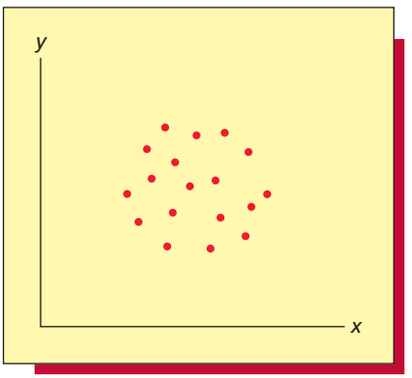
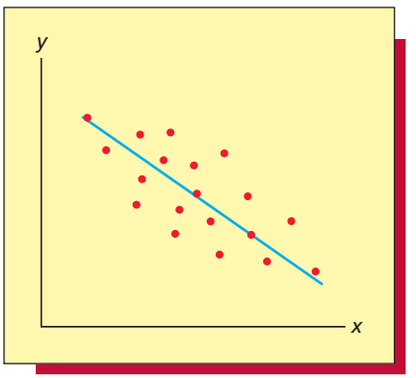


FIGURE 2.27 A Negative Linear Relationship



A scatter plot can reveal various kinds of relationships. For instance, Figures 2.25, 2.26, and 2.27 show several possible relationships between two variables x and y. Figure 2.25 shows a relationship similar to that of our advertising expenditure–sales volume example—here y has a tendency to increase as x increases. Figure 2.26 illustrates a situation in which x and y do not appear to have any linear relationship. Figure 2.27 illustrates a negative linear relationship—here y has a tendency to decrease as x increases. Finally, not all relationships are linear. In Chapter 15 we will consider how to represent and quantify curved relationships.

To conclude this section, recall from Chapter 1 that a **time series plot** (also called a **runs plot**) is a plot of individual process measurements versus time. This implies that a time series plot is a scatter plot, where values of a process variable are plotted on the vertical axis versus corresponding values of time on the horizontal axis.

Exercises for Section 2.6

connect

CONCEPTS

2.52 Explain the purpose for constructing a scatter plot of y versus x.

2.53 Draw a scatter plot of y versus x in which y increases in a linear fashion as x increases.

2.54 Draw a scatter plot of y versus x in which y decreases in a linear fashion as x increases.

2.55 Draw a scatter plot of y versus x in which there is little or no linear relationship between y and x.

2.56 Discuss the relationship between a scatter plot and a time series plot.

METHODS AND APPLICATIONS

2.57 THE REAL ESTATE SALES PRICE CASE DS RealEst

A real estate agency collects data concerning y = the sales price of a house (in thousands of dollars), and x = the home size (in hundreds of square feet). The data are given in Table 2.21. Construct a scatter plot of y versus x and interpret what the plot says.

TABLE 2.21 Real Estate Sales Price Data DS RealEst

Sales Price (y)	Home Size (x)
180	23
98.1	11
173.1	20
136.5	17
141	15
165.9	21
193.5	24
127.8	13
163.5	19
172.5	25

Source: Reprinted with permission from *The Real Estate Appraiser and Analyst,* Spring 1986. Copyright 1986 by the Appraisal Institute, Chicago, Illinois.

2.58 THE NATURAL GAS CONSUMPTION CASE DS GasCon1

Table 2.22 gives the average hourly outdoor temperature (x) in a city during a week and the city's natural gas consumption (y) during the week for each of the previous eight weeks (the temperature readings are expressed in degrees Fahrenheit and the natural gas consumptions are expressed in millions of cubic feet of natural gas). The MINITAB output in Figure 2.28 gives a scatter plot of y versus x. Discuss the nature of the relationship between y and x.

TABLE 2.22 The Natural Gas Consumption Data DS GasCon1

Week	1	2	3	4	5	6	7	8
Temperature, x	28.0	28.0	32.5	39.0	45.9	57.8	58.1	62.5
Natural Gas Consumption, y	12.4	11.7	12.4	10.8	9.4	9.5	8.0	7.5

FIGURE 2.28 MINITAB Scatter Plot of the Natural Gas Consumption Data (for Exercise 2.58)

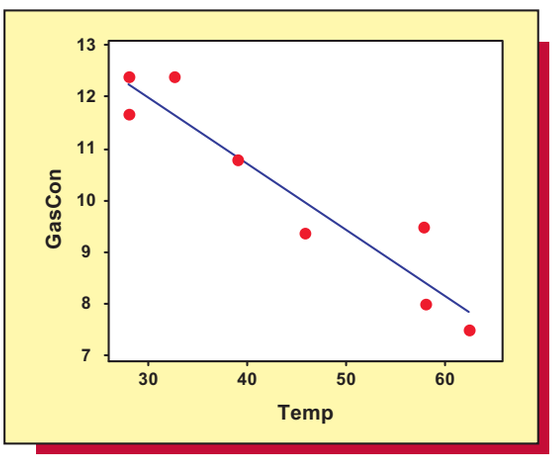


FIGURE 2.29 Time Series Plots for Exercise 2.59 DS PayTVRates

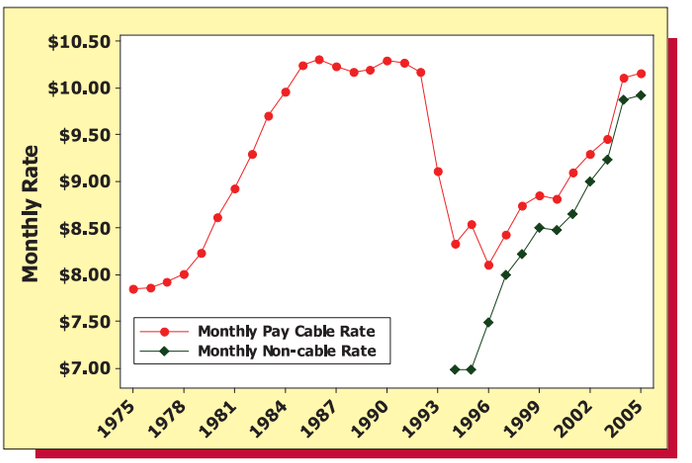
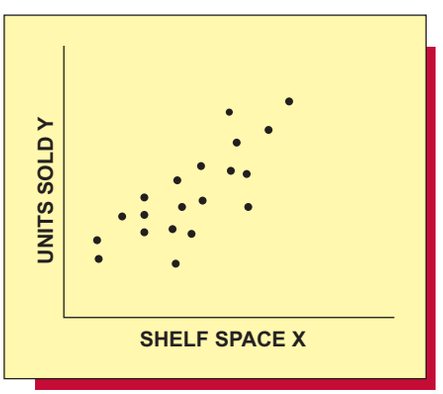


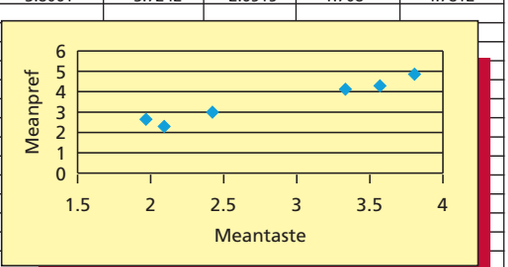
FIGURE 2.30 A Scatter Plot of Units Sold versus Shelf Space (for Exercise 2.60)



Source: W. R. Dillon, T. J. Madden, and N. H. Firtle, *Essentials of Marketing Research* (Burr Ridge, IL: Richard D. Irwin, Inc., 1993), p. 452. Copyright © 1993. Reprinted by permission of McGraw-Hill Companies, Inc.

FIGURE 2.31 Excel Output of the Mean Restaurant Ratings and a Scatter Plot of Mean Preference versus Mean Taste (for Exercise 2.61) DS FastFood

	A	B	C	D	E	F
1	Restaurant	Meantaste	Meanconv	Meanfam	Meanprice	Meanpref
2	Borden Burger	3.5659	2.7005	2.5282	2.9372	4.2552
3	Hardee's	3.329	3.3483	2.7345	2.7513	4.0911
4	Burger King	2.4231	2.7377	2.3368	3.0761	3.0052
5	McDonald's	2.0895	1.938	1.4619	2.4884	2.2429
6	Wendy's	1.9661	2.892	2.3376	4.0814	2.5351
7	White Castle	3.8061	3.7242	2.6515	1.708	4.7812
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						



Source: The Ohio State University.

2.59 Figure 2.29 gives a time series plot of the average U.S. monthly pay cable TV rate (for premium services) for each year from 1975 to 2005. Figure 2.29 also gives a time series plot of the average monthly non-cable (mostly satellite) TV rate (for premium services) for each year from 1994 to 2005.[5] Satellite TV became a serious competitor to cable TV in the early 1990s. Does it appear that the emergence of satellite TV had an influence on cable TV rates? What happened after satellite TV became more established in the marketplace? DS PayTVRates

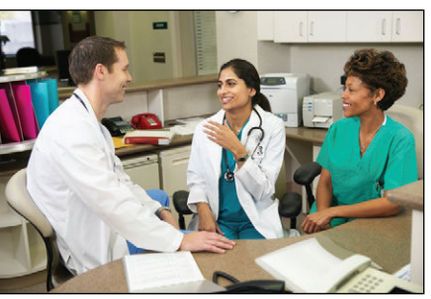
2.60 Figure 2.30 gives a scatter plot of the number of units sold, y, of 20 varieties of a canned soup versus the amount of shelf space, x, allocated to each variety. Do you think that sales are affected by the amount of allocated shelf space, or vice versa?

2.61 THE FAST-FOOD RESTAURANT RATING CASE DS FastFood

Figure 2.31 presents the ratings given by 406 randomly selected individuals of six fast food restaurants on the basis of taste, convenience, familiarity, and price. The data were collected by researchers at The Ohio State University in the early 1990s. Here, 1 is the best rating and 6 the worst. In addition, each individual ranked the restaurants from 1 through 6 on the basis of overall preference. Interpret the Excel scatter plot, and construct and interpret other relevant scatter plots.

[5]The time series data for this exercise are on the website for this book.

LO2-8 Recognize misleading graphs and charts. (Optional)



2.7 Misleading Graphs and Charts (Optional)

The statistical analyst's goal should be to present the most accurate and truthful portrayal of a data set that is possible. Such a presentation allows managers using the analysis to make informed decisions. However, it is possible to construct statistical summaries that are misleading. Although we do not advocate using misleading statistics, you should be aware of some of the ways statistical graphs and charts can be manipulated in order to distort the truth. By knowing what to look for, you can avoid being misled by a (we hope) small number of unscrupulous practitioners.

As an example, suppose that the nurses at a large hospital will soon vote on a proposal to join a union. Both the union organizers and the hospital administration plan to distribute recent salary statistics to the entire nursing staff. Suppose that the mean nurses' salary at the hospital and the mean nurses' salary increase at the hospital (expressed as a percentage) for each of the last four years are as follows:

Year	Mean Salary	Mean Salary Increase (Percent)
1	$60,000	3.0%
2	61,600	4.0
3	63,500	4.5
4	66,100	6.0

The hospital administration does not want the nurses to unionize and, therefore, hopes to convince the nurses that substantial progress has been made to increase salaries without a union. On the other hand, the union organizers wish to portray the salary increases as minimal so that the nurses will feel the need to unionize.

Figure 2.32 gives two bar charts of the mean nurses' salaries at the hospital for each of the last four years. Notice that in Figure 2.32(a) the administration has started the vertical scale of the bar chart at a salary of $58,000 by using a *scale break* (≀). Alternatively, the chart could be set up without the scale break by simply starting the vertical scale at $58,000. Starting the vertical scale at a value far above zero makes the salary increases look more dramatic. Notice that when the union organizers present the bar chart in Figure 2.32(b), which has a vertical scale starting at zero, the salary increases look far less impressive.

Figure 2.33 presents two bar charts of the mean nurses' salary increases (in percentages) at the hospital for each of the last four years. In Figure 2.33(a), the administration has made the widths of the bars representing the percentage increases proportional to their heights. This makes the upward movement in the mean salary increases look more dramatic because the observer's eye tends to compare the areas of the bars, while the improvements in the mean salary increases are really only proportional to the heights of the bars. When the union organizers present the bar chart of Figure 2.33(b), the improvements in the mean salary increases look less impressive because each bar has the same width.

FIGURE 2.32 Two Bar Charts of the Mean Nurses' Salaries at a Large Hospital for the Last Four Years

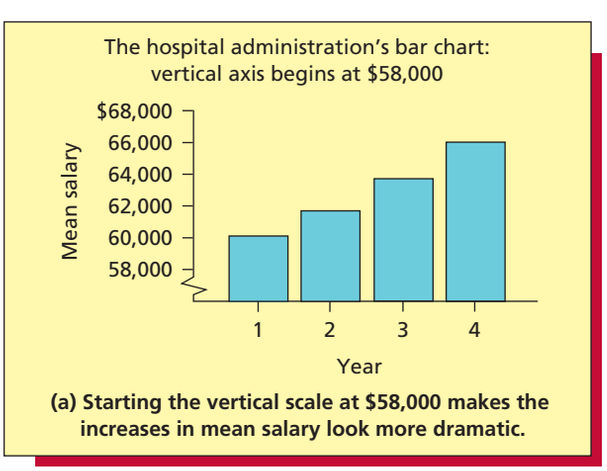
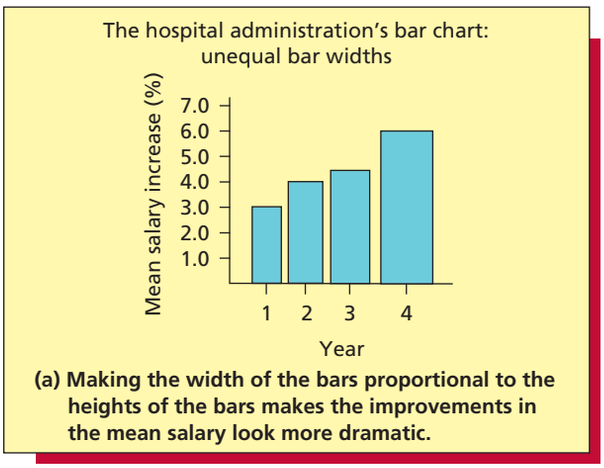
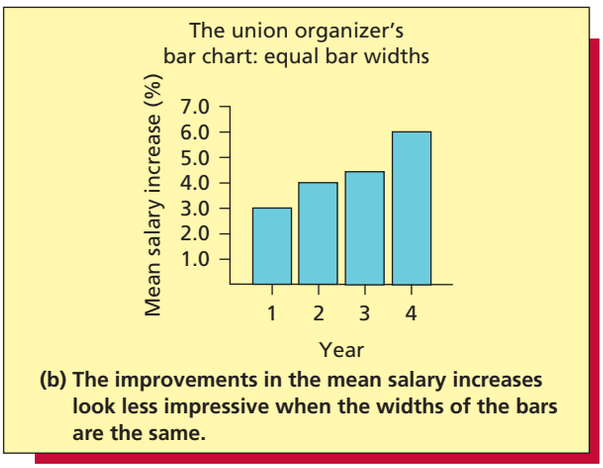


FIGURE 2.33 Two Bar Charts of the Mean Nurses' Salary Increases at a Large Hospital for the Last Four Years.

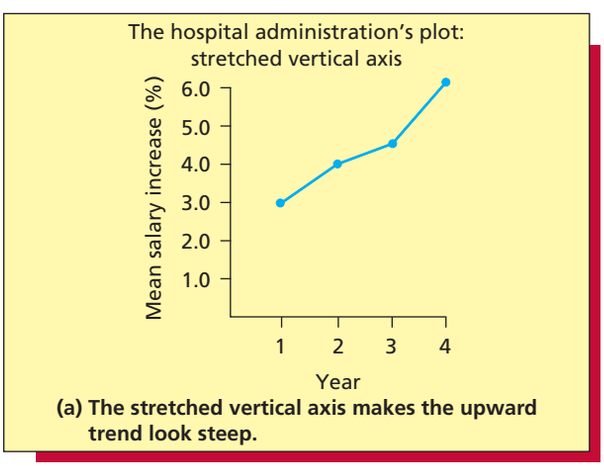


(a) Making the width of the bars proportional to the heights of the bars makes the improvements in the mean salary look more dramatic.

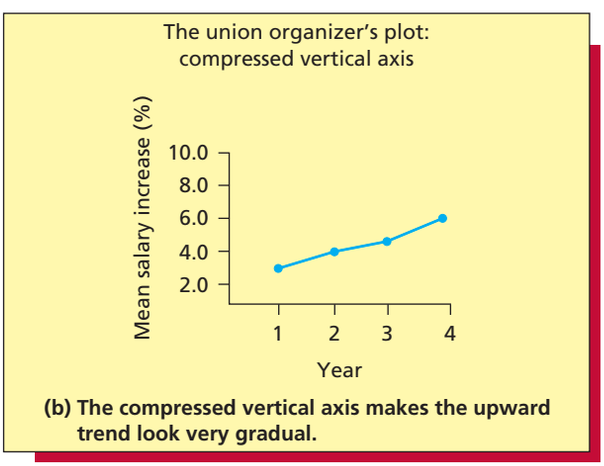


(b) The improvements in the mean salary increases look less impressive when the widths of the bars are the same.

FIGURE 2.34 Two Time Series Plots of the Mean Nurses' Salary Increases at a Large Hospital for the Last Four Years



(a) The stretched vertical axis makes the upward trend look steep.



(b) The compressed vertical axis makes the upward trend look very gradual.

Figure 2.34 gives two time series plots of the mean nurses' salary increases at the hospital for the last four years. In Figure 2.34(a) the administration has stretched the vertical axis of the graph. That is, the vertical axis is set up so that the distances between the percentages are large. This makes the upward trend of the mean salary increases appear to be steep. In Figure 2.34(b) the union organizers have compressed the vertical axis (that is, the distances between the percentages are small). This makes the upward trend of the mean salary increases appear to be gradual. As we will see in the exercises, stretching and compressing the horizontal axis in a time series plot can also greatly affect the impression given by the plot.

It is also possible to create totally different interpretations of the same statistical summary by simply using different labeling or captions. For example, consider the bar chart of mean nurses' salary increases in Figure 2.33(b). To create a favorable interpretation, the hospital administration might use the caption "Salary Increase Is Higher for the Fourth Year in a Row." On the other hand, the union organizers might create a negative impression by using the caption "Salary Increase Fails to Reach 10% for Fourth Straight Year."

In summary, we do not approve of using statistics to mislead and distort reality. Statistics should be used to present the most truthful and informative summary of the data that is possible. However, it is important to carefully study any statistical summary so that you will not be misled. Look for manipulations such as stretched or compressed axes on graphs, axes that do not begin at zero, and bar charts with bars of varying widths. Also, carefully think about assumptions, and make your own conclusions about the meaning of any statistical summary rather than relying on captions written by others. Doing these things will help you to see the truth and to make well-informed decisions.

Exercises for Section 2.7

CONCEPTS

connect

2.62 When we construct a bar chart or graph, what is the effect of starting the vertical axis at a value that is far above zero? Explain.

2.63 Find an example of a misleading use of statistics in a newspaper, magazine, corporate annual report, or other source. Then explain why your example is misleading.

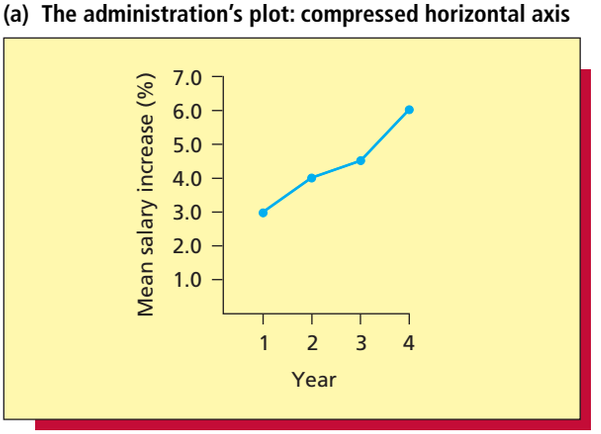
METHODS AND APPLICATIONS

2.64 Figure 2.35 gives two more time series plots of the previously discussed mean nurses' salary increases. In Figure 2.35(a) the hospital administration has compressed the horizontal axis. In Figure 2.35(b) the union organizers have stretched the horizontal axis. Discuss the different impressions given by the two time series plots.

2.65 In the article "How to Display Data Badly" in the May 1984 issue of *The American Statistician,* Howard Wainer presents a *stacked bar chart* of the number of public and private elementary schools (1929–1970). This bar chart is given in Figure 2.36. Wainer also gives a line graph of the number of private elementary schools (1930–1970). This graph is shown in Figure 2.37.

FIGURE 2.35 Two Time Series Plots of the Mean Nurses' Salary Increases at a Large Hospital for the Last Four Years (for Exercise 2.64)

(a) The administration's plot: compressed horizontal axis



(b) The union organizer's plot: stretched horizontal axis

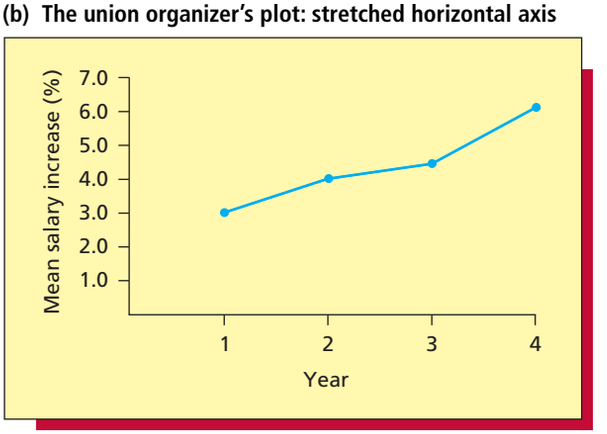


FIGURE 2.36 Wainer's Stacked Bar Chart (for Exercise 2.65)

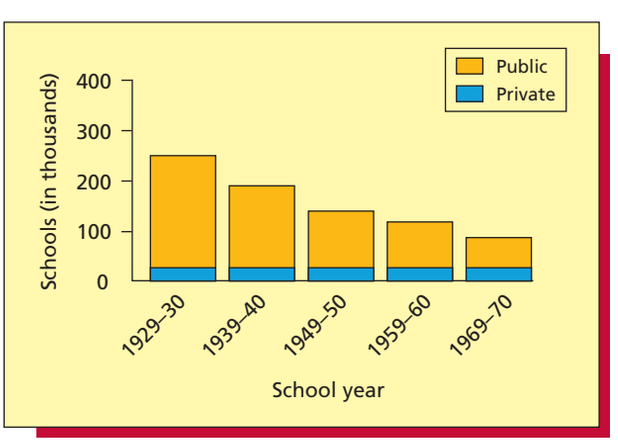
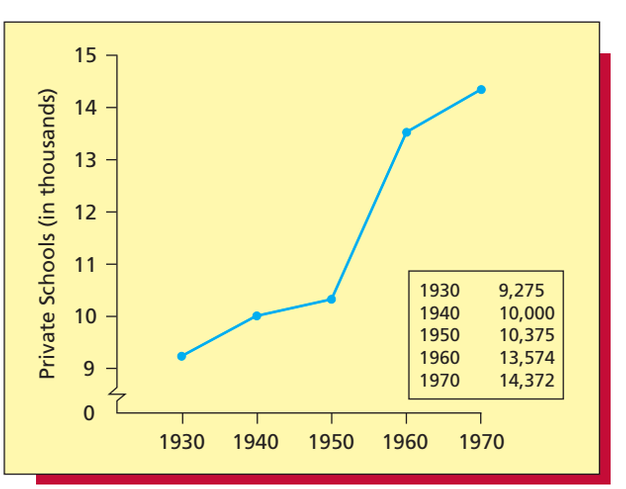


FIGURE 2.37 Wainer's Line Graph (for Exercise 2.65)



Source: H. Wainer, "How to Display Data Badly," *The American Statistician,* May 1984, pp. 137–147. Copyright © 1984 American Statistical Association. Used with permission.

a Looking at the bar chart of Figure 2.36, does there appear to be an increasing trend in the number of private elementary schools from 1930 to 1970?
b Looking at the line graph of Figure 2.37, does there appear to be an increasing trend in the number of private elementary schools from 1930 to 1970?
c Which portrayal of the data do you think is more appropriate? Explain why.
d Is either portrayal of the data entirely appropriate? Explain.

Chapter Summary

We began this chapter by explaining how to summarize qualitative data. We learned that we often summarize this type of data in a table that is called a **frequency distribution.** Such a table gives the **frequency, relative frequency,** or **percent frequency** of items that are contained in each of several nonoverlapping classes or categories. We also learned that we can summarize qualitative data in graphical form by using **bar charts** and **pie charts** and that qualitative quality data are often summarized using a special bar chart called a **Pareto chart.** We continued in Section 2.2 by discussing how to graphically portray quantitative data. In particular, we explained how to summarize such data by using frequency distributions and histograms. We saw that a **histogram** can be constructed using frequencies, relative frequencies, or percentages, and that we often construct histograms using statistical software such as MINITAB or the analysis toolpak in Excel. We used histograms to describe the shape of a distribution and we saw that distributions are sometimes **mound shaped and symmetrical,** but that a distribution can also be **skewed** (**to the right** or **to the left**). We also learned that a frequency distribution can be graphed by using a **frequency polygon** and that a graph of a **cumulative frequency distribution** is called an **ogive.** In Sections 2.3 and 2.4 we showed how to summarize relatively small data sets by using **dot plots** and **stem-and-leaf displays.** These graphics allow us to see all of the measurements in a data set and to (simultaneously) see the shape of the data set's distribution. Next, we learned about how to describe the relationship between two variables. First, in optional Section 2.5 we explained how to construct and interpret a **contingency table,** which classifies data on two dimensions using a table that consists of rows and columns. Then, in optional Section 2.6 we showed how to construct a **scatter plot.** Here, we plot numerical values of one variable on a horizontal axis versus numerical values of another variable on a vertical axis. We saw that we often use such a plot to look at possible straight-line relationships between the variables. Finally, in optional Section 2.7 we learned about misleading graphs and charts. In particular, we pointed out several graphical tricks to watch for. By careful analysis of a graph or chart, one can avoid being misled.

Glossary of Terms

bar chart: A graphical display of data in categories made up of vertical or horizontal bars. Each bar gives the frequency, relative frequency, or percentage frequency of items in its corresponding category. (page 36)
class midpoint: The point in a class that is halfway between the lower and upper class boundaries. (page 45)
contingency table: A table consisting of rows and columns that is used to classify data on two dimensions. (page 61)
cumulative frequency distribution: A table that summarizes the number of measurements that are less than the upper class boundary of each class. (page 49)
cumulative percent frequency distribution: A table that summarizes the percentage of measurements that are less than the upper class boundary of each class. (page 50)
cumulative relative frequency distribution: A table that summarizes the fraction of measurements that are less than the upper class boundary of each class. (page 50)
dot plot: A graphical portrayal of a data set that shows the data set's distribution by plotting individual measurements above a horizontal axis. (page 54)
frequency distribution: A table that summarizes the number of items (or measurements) in each of several nonoverlapping classes. (pages 35, 44)
frequency polygon: A graphical display in which we plot points representing each class frequency (or relative frequency or percent frequency) above their corresponding class midpoints and connect the points with line segments. (page 48)
histogram: A graphical display of a frequency distribution, relative frequency distribution, or percentage frequency distribution. It divides measurements into classes and graphs the frequency, relative frequency, or percentage frequency for each class. (pages 42, 44)
ogive: A graph of a cumulative distribution (frequencies, relative frequencies, or percent frequencies may be used). (page 50)
outlier: An unusually large or small observation that is well separated from the remaining observations. (pages 55, 59)
Pareto chart: A bar chart of the frequencies or percentages for various types of defects. These are used to identify opportunities for improvement. (page 38)
percent frequency distribution: A table that summarizes the percentage of items (or measurements) in each of several nonoverlapping classes. (pages 36, 44)
pie chart: A graphical display of data in categories made up of "pie slices." Each pie slice represents the frequency, relative frequency, or percentage frequency of items in its corresponding category. (page 37)
relative frequency distribution: A table that summarizes the fraction of items (or measurements) in each of several nonoverlapping classes. (pages 36, 44)
scatter plot: A graph that is used to study the possible relationship between two variables y and x. The observed values of y are

plotted on the vertical axis versus corresponding observed values of x on the horizontal axis. (page 67)

skewed to the left: A distribution shape having a long tail to the left. (page 47)

skewed to the right: A distribution shape having a long tail to the right. (page 47)

stem-and-leaf display: A graphical portrayal of a data set that shows the data set's distribution by using stems consisting of leading digits and leaves consisting of trailing digits. (page 56)

symmetrical distribution: A distribution shape having right and left sides that are "mirror images" of each other. (page 47)

Important Formulas and Graphics

Frequency distribution: pages 35, 44, 45

Relative frequency: pages 36, 45

Percent frequency: pages 36, 45

Bar chart: pages 36, 37

Pie chart: page 37

Pareto chart: page 38

Histogram: page 45

Frequency polygon: page 48

Cumulative distribution: page 49

Ogive: page 50

Dot plot: page 54

Stem-and-leaf display: page 58

Contingency table: page 61

Scatter plot: page 67

Time series plot: page 68

Supplementary Exercises

connect

2.66 At the end of 2011 Chrysler Motors was trying to decide whether or not to discontinue production of a Jeep model introduced in 2002—the Jeep Liberty. Although Liberty sales had generally declined since 2007 with the cancellation of one Jeep model and the introduction of three new Jeep models, Liberty sales had continued to be good (and the Liberty name had retained a positive image) at some Jeep dealerships. Suppose that the owner of one such dealership in Cincinnati, Ohio, wished to present Chrysler with evidence that Liberty sales were still an important part of total sales at his dealership. To do this, the owner decided to compare his dealership's Liberty sales in 2011 with those in 2006. Recently summarized sales data show that in 2011 the owner's dealership sold 30 Jeep Compasses, 50 Jeep Grand Cherokees, 45 Jeep Libertys, 40 Jeep Patriots, 28 Jeep Wranglers, and 35 Jeep Wrangler Unlimiteds. Such summarized sales data were not available for 2006, but raw sales data were found. Denoting the four Jeep models sold in 2006 (Commander, Grand Cherokee, Liberty, and Wrangler) as C, G, L, and W, these raw sales data are shown in Table 2.23. Construct percent bar charts of (1) Jeep sales in 2011 and (2) Jeep sales in 2006. Compare the charts and write a short report supporting the owner's position that in 2011 Liberty sales were still an important part of sales at his dealership. DS JeepSales

TABLE 2.23 2006 Sales at a Greater Cincinnati Jeep Dealership DS JeepSales

W	L	L	W	G	C	C	L	C	L	G	W	C	L	L	C	C	G	L	L	W	C	G	G	C	L
L	L	G	L	C	C	G	C	C	G	C	L	W	W	G	G	W	G	C	W	W	G	L	L	G	
G	L	G	C	C	C	C	C	G	G	L	G	G	L	L	L	C	C	G	C	L	G	G	G	L	
L	G	L	L	G	L	C	W	G	L	G	L	G	G	G	C	G	W	G	L	L	L	C	C	L	
G	L	C	L	C	L	L	L	C	G	L	C	L	W	L	W	G	W	C	W	C	W	C	L	C	
C	G	C	C	C	C	C	C	C	G	C	C	W	G	C	G	L	L	C	L	L	G	G	G	L	
L	L	C	G	L	C	C	L	L	G	G	L	L	L	C	G	L	C	L	W	L	L	C	G	C	
G	G	G	L	C	L	L	G	L	C	C	L	G	W	W	W	C	C	C	G	G	L	G	C	G	
C	L	L	G	G	L	W	W	L	C	C	C	G	W	C	C	W	L	G	W	L	L	L	G	G	
G	W	L	L	C	G	C	C	W	C	L	L	L	G	G	W	L	L	C	L	G	G	W	G	G	

Exercises 2.67 through 2.74 are based on the data in Table 2.24. This table gives the results of the J.D. Power initial quality study of 2010 automobiles. Each model is rated on overall mechanical quality and overall design quality on a scale from "among the best" to "the rest"—see the scoring legend at the bottom of the table. DS JDPower

TABLE 2.24 Results of the J. D. Power Initial Quality Study of 2010 Automobiles JDPower

Company	Country of Origin	Overall Quality Mechanical	Overall Quality Design
Acura	Japan	●●●●	●●●●
Audi	Germany	●●●	●●●
BMW	Germany	●●●	●●●
Buick	United States	●●●	●●
Cadillac	United States	●●●	●●●
Chevrolet	United States	●●●	●●●
Chrysler	United States	●●●	●●●
Dodge	United States	●●●	●●
Ford	United States	●●●	●●●●
GMC	United States	●●●	●●
Honda	Japan	●●●	●●●●
Hyundai	Korea	●●●	●●●●
Infiniti	Japan	●●●	●●●●
Jaguar	Great Britain	●●	●●●
Jeep	United States	●●	●●●
Kia	Korea	●●	●●●
Land Rover	Great Britain	●●	●●●
Lexus	Japan	●●●●●	●●●
Lincoln	United States	●●●	●●●
Mazda	Japan	●●●	●●●
Mercedes-Benz	Germany	●●●●●	●●●
Mercury	United States	●●●	●●●
MINI	Great Britain	●●●	●●
Mitsubishi	Japan	●●	●●●
Nissan	Japan	●●●	●●●
Porsche	Germany	●●●●	●●●●●
Ram	United States	●●●	●●●●
Scion	Japan	●●●	●●●
Subaru	Japan	●●●	●●●
Suzuki	Japan	●●●	●●●
Toyota	Japan	●●●	●●●
Volkswagen	Germany	●●	●●●
Volvo	Sweden	●●●	●●●

Scoring Legend

●●●●● Among the best ●●● About average

●●●● Better than most ●● The rest

Source: http://www.jdpower.com/autos/quality-ratings-by-brand/

2.67 Develop a frequency distribution of the overall mechanical quality ratings. Describe the distribution. JDPower

2.68 Develop a relative frequency distribution of the overall design quality ratings. Describe the distribution. JDPower

2.69 Construct a percentage bar chart of the overall mechanical quality ratings for each of the following: automobiles of United States origin; automobiles of Pacific Rim origin (Japan/Korea); and automobiles of European origin (Germany/Great Britain/Sweden). Compare the three distributions in a written report. JDPower

2.70 Construct a percentage pie chart of the overall design quality ratings for each of the following: automobiles of United States origin; automobiles of Pacific Rim origin (Japan/Korea); and automobiles of European origin (Germany/Great Britain/Sweden). Compare the three distributions in a written report. JDPower

2.71 Construct a contingency table of automobile origin versus overall mechanical quality rating. Set up rows corresponding to the United States, the Pacific Rim (Japan/Korea), and Europe (Germany/Great Britain/Sweden), and set up columns corresponding to the ratings "among the best" through "the rest." Describe any apparent relationship between origin and overall mechanical quality rating. JDPower

2.72 Develop a table of row percentages for the contingency table you set up in Exercise 2.71. Using these row percentages, construct a percentage frequency distribution of overall mechanical quality rating for each of the United States, the Pacific Rim, and Europe. Illustrate these three frequency distributions using percent bar charts and compare the distributions in a written report. JDPower

2.73 Construct a contingency table of automobile origin versus overall design quality rating. Set up rows corresponding to the United States, the Pacific Rim (Japan/Korea), and Europe (Germany/Great Britain/Sweden), and set up columns corresponding to the ratings "among the best" through "the rest." Describe any apparent relationship between origin and overall design quality. JDPower

2.74 Develop a table of row percentages for the contingency table you set up in Exercise 2.73. Using these row percentages, construct a percentage frequency distribution of overall design quality rating for each of the United States, the Pacific Rim, and Europe. Illustrate these three frequency distributions using percentage pie charts and compare the distributions in a written report. JDPower

THE CIGARETTE ADVERTISEMENT CASE DS ModelAge

In an article in the *Journal of Marketing,* Mazis, Ringold, Perry, and Denman discuss the perceived ages of models in cigarette advertisements.[6] To quote the authors:

> Most relevant to our study is the Cigarette Advertiser's Code, initiated by the tobacco industry in 1964. The code contains nine advertising principles related to young people, including the following provision (*Advertising Age* 1964): "Natural persons depicted as smokers in cigarette advertising shall be at least 25 years of age and shall not be dressed or otherwise made to appear to be less than 25 years of age."

Tobacco industry representatives have steadfastly maintained that code provisions are still being observed. A 1988 Tobacco Institute publication, "Three Decades of Initiatives by a Responsible Cigarette Industry," refers to the industry code as prohibiting advertising and promotion "directed at young people" and as "requiring that models in advertising must be, and must appear to be, at least 25 years old." John R. Nelson, Vice President of Corporate Affairs for Philip Morris, wrote, "We employ only adult models in our advertising who not only are but *look* over 25." However, industry critics have charged that current cigarette advertising campaigns use unusually young-looking models, thereby violating the voluntary industry code.

Suppose that a sample of 50 people is randomly selected at a shopping mall. Each person in the sample is shown a typical cigarette advertisement and is asked to estimate the age of the model in the ad. The 50 perceived age estimates so obtained are given below. Use these data to do Exercises 2.75 through 2.78.

26	30	23	27	27	32	28	19	25	29
31	28	24	26	29	27	28	17	28	21
30	28	25	31	22	29	18	27	29	23
28	26	24	30	27	25	26	28	20	24
29	32	27	17	30	27	21	29	26	28

2.75 Consider constructing a frequency distribution and histogram for the perceived age estimates. DS ModelAge

- **a** Develop a frequency distribution, a relative frequency distribution, and a percent frequency distribution for the perceived age estimates using eight classes each of length 2. Note that, while the procedure presented in Section 2.2 would tell us to use six classes, in this case we get a more informative frequency distribution by using eight classes.
- **b** Draw a frequency histogram for the perceived age estimates.
- **c** Describe the shape of the distribution of perceived age estimates.

2.76 Construct a frequency polygon of the perceived age estimates. Use the classes of Exercise 2.75. DS ModelAge

2.77 Construct a dot plot of the perceived age estimates and describe the shape of the distribution. What percentage of the perceived ages are below the industry's code provision of 25 years old? Do you think that this percentage is too high? DS ModelAge

2.78 Using the frequency distribution you developed in Exercise 2.75, develop: DS ModelAge

- **a** A cumulative frequency distribution.
- **b** A cumulative relative frequency distribution.
- **c** A cumulative percent frequency distribution.
- **d** A frequency ogive of the perceived age estimates.
- **e** How many perceived age estimates are 28 or less?
- **f** What percentage of perceived age estimates are 22 or less?

2.79 Table 2.25 presents data concerning the largest U.S. charities as rated on the Forbes.com website on June 14, 2011.

- **a** Construct a percent frequency histogram of each of (1) the charities' private support figures, (2) the charities' total revenues, and (3) the charities' fundraising efficiencies.
- **b** Describe the shape of each histogram. DS Charities

2.80 The price/earnings ratio of a firm is a multiplier applied to a firm's earnings per share (EPS) to determine the value of the firm's common stock. For instance, if a firm's earnings per share is $5,

[6] Source: M. B. Mazis, D. J. Ringold, E. S. Perry, and D. W. Denman, "Perceived Age and Attractiveness of Models in Cigarette Advertisements," *Journal of Marketing* 56 (January 1992), pp. 22–37.

TABLE 2.25 Data Concerning the Largest U.S. Charities as Rated by *Forbes* Magazine DS Charities

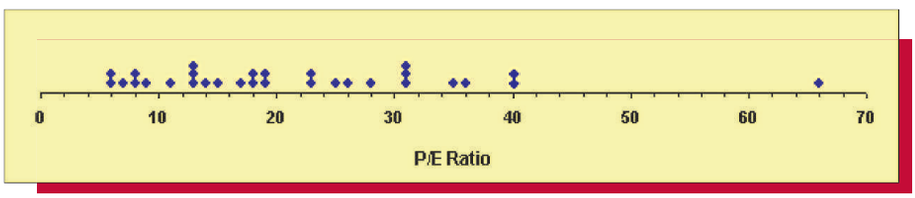
Name	Private Support ($mil)	Total Revenue ($mil)	Fundraising Efficiency (%)
Academy for Educational Development	51	414	100
Aeras Global TB Vaccine Foundation	63	71	100
Alzheimer's Association	189	206	80
American Cancer Society	898	897	75
American Civil Liberties Foundation	62	73	88
American Diabetes Association	156	205	70
American Enterprise Institute	59	51	98
American Heart Association	440	464	75
American Jewish Joint Distribution Committee	245	244	99
American Kidney Foundation	173	172	98
American Museum of Natural History	63	135	91
American National Red Cross	661	2.95	81
American Nicaraguan Foundation	193	197	100
American SPCA	91	139	78
AmeriCares Foundation	1.19	1.19	99
Anti-Defamation League of B'nai B'rith	49	66	84
Art Institute of Chicago	77	180	89
Arthritis Foundation	84	117	85
Asia Foundation	51	143	99
Big Brothers Big Sisters of America	214	295	83
Billy Graham Evangelistic Association	85	96	94
Boston Symphony Orchestra	52	100	85
Boy Scouts of America	375	1.13	87
Boys & Girls Clubs of America	625	1.47	88
Brookings Institute	57	75	95
Brother's Brother Foundation	267	269	100
Campus Crusade for Christ	473	509	92
Cape Cod Healthcare	57	603	95
CARE USA	384	692	94
Carter Center	153	114	95
Catholic Charities USA	939	4.27	93
Catholic Medical Mission Board	279	280	99
Catholic Relief Services	220	777	90
ChildFund International	216	211	90
Children International	146	146	88
Children's Hospital Boston	89	1.37	76
Children's Hospital Los Angeles	95	547	86
Children's Hospital Philadelphia	52	1.87	73
Children's Memorial Hospital	68	624	75
Children's Network International	62	62	100

Source: http://www.forbes.com/lists/2010/14/charity-10_land.html (accessed June 14, 2011).

and if its price/earnings ratio (or P/E ratio) is 10, then the market value of each share of common stock is ($5)(10) = $50. To quote Stanley B. Block and Geoffrey A. Hirt in their book *Foundations of Financial Management*:[7]

> The P/E ratio indicates expectations about the future of a company. Firms expected to provide returns greater than those for the market in general with equal or less risk often have P/E ratios higher than the market P/E ratio.

In the figure below we give a dot plot of the P/E ratios for 30 fast-growing companies. Describe the distribution of the P/E ratios.



2.81 In this exercise we consider how to deal with class lengths that are unequal (and with open-ended classes) when setting up histograms. Often data are published in this form and we wish to

[7]Source: Excerpt from S. B. Block and G. A. Hirt, *Foundations of Financial Management,* p. 28. © 1994 Richard D. Irwin. Reprinted with permission of McGraw-Hill Companies, Inc.

DS ISO 9000

Annual Savings	Number of Companies
0 to $10K	162
$10K to 25K	62
$25K to 50K	53
$50K to 100K	60
$100K to 150K	24
$150K to 200K	19
$200K to 250K	22
$250K to 500K	21
(>$500K)	37

Note: (K = 1000)

construct a histogram. An example is provided by data concerning the benefits of ISO 9000 registration published by CEEM Information Services. According to CEEM:[8]

> ISO 9000 is a series of international standards for quality assurance management systems. It establishes the organizational structure and processes for assuring that the production of goods or services meet a consistent and agreed-upon level of quality for a company's customers.

CEEM presents the results of a Quality Systems Update/Deloitte & Touche survey of ISO 9000–registered companies conducted in July 1993. Included in the results is a summary of the total annual savings associated with ISO 9000 implementation for surveyed companies. The findings (in the form of a frequency distribution of ISO 9000 savings) are given on the page margin. Notice that the classes in this distribution have unequal lengths and that there is an open-ended class (>$500K).

To construct a histogram for these data, we select one of the classes as a base. It is often convenient to choose the shortest class as the base (although it is not necessary to do so). Using this choice, the 0 to $10K class is the base. This means that we will draw a rectangle over the 0 to $10K class having a height equal to 162 (the frequency given for this class in the published data). Because the other classes are longer than the base, the heights of the rectangles above these classes will be adjusted. To do this we employ a rule that says that the area of a rectangle positioned over a particular class should represent the relative proportion of measurements in the class. Therefore, we proceed as follows. The length of the $10K to 25K class differs from the base class by a factor of $(25 - 10)/(10 - 0) = 3/2$, and, therefore, we make the height of the rectangle over the $10K to 25K class equal to $(2/3)(62) = 41.333$. Similarly, the length of the $25K to 50K class differs from the length of the base class by a factor of $(50 - 25)/(10 - 0) = 5/2$, and, therefore, we make the height of the rectangle over the $25K to 50K class equal to $(2/5)(53) = 21.2$.

a Use the procedure just outlined to find the heights of the rectangles drawn over all the other classes (with the exception of the open-ended class, >$500K).

b Draw the appropriate rectangles over the classes (except for >$500K). Note that the $250K to 500K class is a lot longer than the others. There is nothing wrong with this as long as we adjust its rectangle's height.

c We complete the histogram by placing a star (∗) to the right of $500K on the scale of measurements and by noting "37" next to the ∗ to indicate 37 companies saved more than $500K. Complete the histogram by doing this.

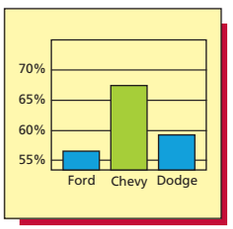
2.82 A basketball player practices free throws by taking 25 shots each day, and he records the number of shots missed each day in order to track his progress. The numbers of shots missed on days 1 through 30 are, respectively, 17, 15, 16, 18, 14, 15, 13, 12, 10, 11, 11, 10, 9, 10, 9, 9, 9, 10, 8, 10, 6, 8, 9, 8, 7, 9, 8, 7, 5, 8. DS FreeThrw

a Construct a stem-and-leaf display and a time series plot of the numbers of missed shots.

b Do you think that the stem-and-leaf display is representative of the numbers of shots that the player will miss on future days? Why or why not?

2.83 Figure 2.38 was used in various Chevrolet magazine advertisements in 1997 to compare the overall resale values of Chevrolet, Dodge, and Ford trucks in the years from 1990 to 1997. What is somewhat misleading about this graph?

FIGURE 2.38
A Graph Comparing the Resale Values of Chevy, Dodge, and Ford Trucks



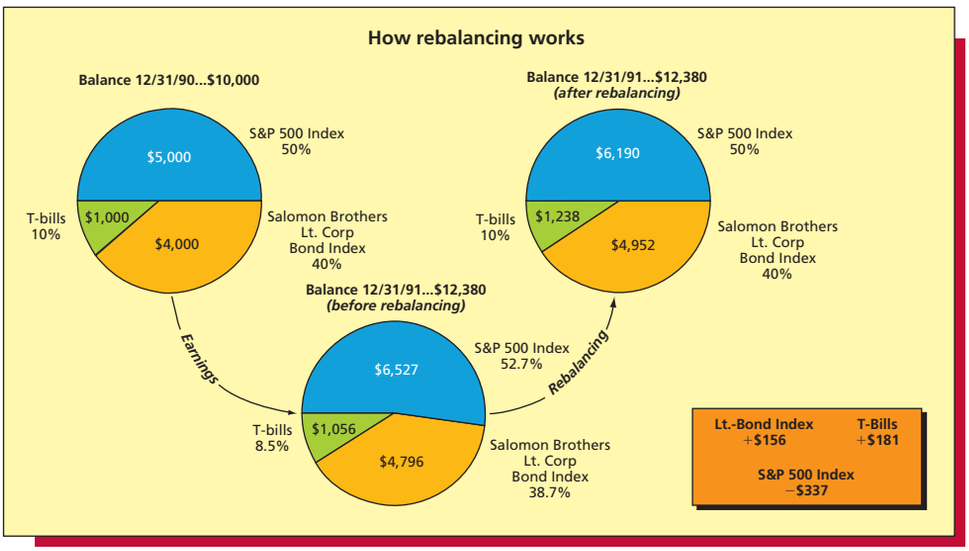
Source: Reprinted courtesy of General Motors Corporation.

2.84 In the Fall 1993 issue of *VALIC Investment Digest,* the Variable Annuity Life Insurance Company used pie charts to illustrate an investment strategy called **rebalancing.** This strategy involves reviewing an investment portfolio annually to return the asset mix (stocks, bonds, Treasury bills, and so on) to a preselected allocation mix. VALIC describes rebalancing as follows (refer to the pie charts in Figure 2.39):

> *Rebalancing—A Strategy to Keep Your Allocation on Track*
>
> Once you've established your ideal asset allocation mix, many experts recommend that you review your portfolio at least once a year to make sure your portfolio remains consistent with your preselected asset allocation mix. This practice is referred to as *rebalancing.*
>
> For example, let's assume a moderate asset allocation mix of 50 percent equities funds, 40 percent bond funds, and 10 percent cash-equivalent funds. The chart [see Figure 2.39] based on data provided by Ibbotson, a major investment and consulting firm, illustrates how rebalancing works. Using the Standard & Poor's 500 Index, the Salomon Brothers Long-Term High-Grade Corporate Bond Index, and the U.S. 30-day Treasury bill average as a cash-equivalent rate, our hypothetical portfolio balance on 12/31/90 is $10,000. One year later the account had grown to $12,380. By the end of 1991, the allocation had changed to 52.7%/38.7%/8.5%. The third pie chart illustrates how the account was once again rebalanced to return to a 50%/40%/10% asset allocation mix.

[8] Source: CEEM Information Services, Fairfax, Virginia. *Is ISO 9000 for You?*

FIGURE 2.39 Using Pie Charts to Illustrate Portfolio Rebalancing (for Exercise 2.84)



Source: The Variable Annuity Life Insurance Company, *VALIC* 6, no. 4 (Fall 1993).

Rebalancing has the potential for more than merely helping diversify your portfolio. By continually returning to your original asset allocation, it is possible to avoid exposure to more risk than you previously decided you were willing to assume.

a Suppose you control a $100,000 portfolio and have decided to maintain an asset allocation mix of 60 percent stock funds, 30 percent bond funds, and 10 percent government securities. Draw a pie chart illustrating your portfolio (like the ones in Figure 2.39).

b Over the next year your stock funds earn a return of 30 percent, your bond funds earn a return of 15 percent, and your government securities earn a return of 6 percent. Calculate the end-of-year values of your stock funds, bond funds, and government securities. After calculating the end-of-year value of your entire portfolio, determine the asset allocation mix (percent stock funds, percent bond funds, and percent government securities) of your portfolio before rebalancing. Finally, draw an end-of-year pie chart of your portfolio before rebalancing.

c Rebalance your portfolio. That is, determine how much of the portfolio's end-of-year value must be invested in stock funds, bond funds, and government securities in order to restore your original asset allocation mix of 60 percent stock funds, 30 percent bond funds, and 10 percent government securities. Draw a pie chart of your portfolio after rebalancing.

2.85 Internet Exercise

The Gallup Organization provides market research and consulting services around the world. Gallup publishes the Gallup Poll, a widely recognized barometer of American and international opinion. The Gallup website provides access to many recent Gallup studies. Although a subscription is needed to access the entire site, many articles about recent Gallup Poll results can be accessed free of charge. To find poll results, go to the Gallup home page (http://www.gallup.com/) and click on the Gallup Poll icon or type the web address http://www.galluppoll.com/ directly into your web browser. The poll results are presented using a variety of statistical summaries and graphics that we have learned about in this chapter.

a Go to the Gallup Organization website and access several of the articles presenting recent poll results. Find and print examples of some of the statistical summaries and graphics that we studied in this chapter. Then write a summary describing which statistical methods and graphics seem to be used most frequently by Gallup when presenting poll results.

b Read the results of a Gallup poll that you find to be of particular interest and summarize (in your own words) its most important conclusions. Cite the statistical evidence in the article that you believe most clearly backs up each conclusion.

c By searching the web, or by searching other sources (such as newspapers and magazines), find an example of a misleading statistical summary or graphic. Print or copy the misleading example and write a paragraph describing why you think the summary or graphic is misleading.

Appendix 2.1 ■ Tabular and Graphical Methods Using Excel

The instructions in this section begin by describing the entry of data into an Excel spreadsheet. Alternatively, the data may be downloaded from this book's website. The appropriate data file name is given at the top of each instruction block. Please refer to Appendix 1.1 for further information about entering data, saving data, and printing results in Excel.

Construct a frequency distribution and frequency bar chart of pizza preferences as in Table 2.2 and Figure 2.1 on pages 36 and 37 (data file: PizzaPref.xlsx):

- Enter the pizza preference data in Table 2.1 on page 36 into column A with label Preference in cell A1.

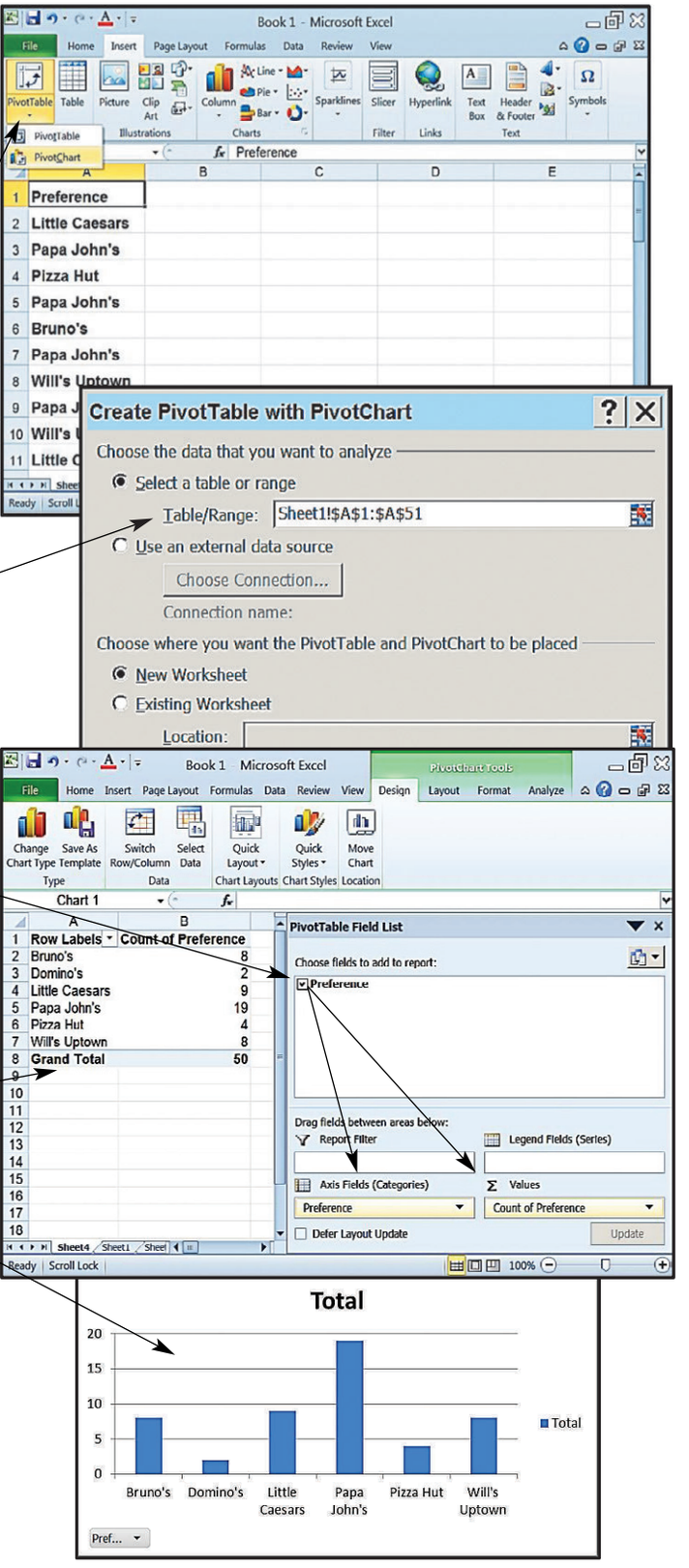
We obtain the frequency distribution and bar chart by forming what is called a **PivotTable**. This is done as follows:

- Select **Insert : PivotTable : PivotChart**

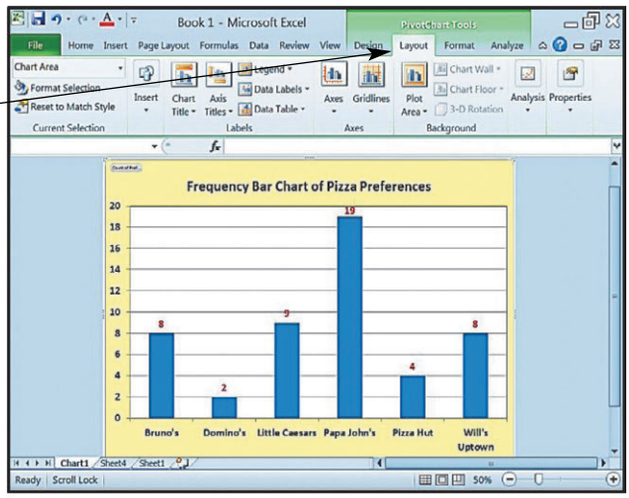
  Note: Be sure to click on the arrow below PivotTable in order to obtain the PivotChart selection.
- In the Create PivotTable with PivotChart dialog box, click "Select a table or range."
- Enter the range of the data to be analyzed into the Table/Range window. Here we have entered the range of the pizza preference data A1:A51—that is, the entries in rows 1 through 51 in column A. The easiest way to do this is to click in the Table/Range window and to then use the mouse to drag the cursor from cell A1 through cell A51.
- Select "New Worksheet" to have the PivotTable and PivotChart output displayed in a new worksheet.
- Click OK in the Create PivotTable with PivotChart dialog box.
- In the **PivotTable Field List task pane,** place a checkmark in the checkbox to the left of the column label "Preference"—when you do this, the label "Preference" will also be placed in the Axis Fields area.
- Also drag the label "Preference" and drop it into the Σ Values area. When this is done, the label will automatically change to "Count of Preference" and the PivotTable will be displayed in the new worksheet.

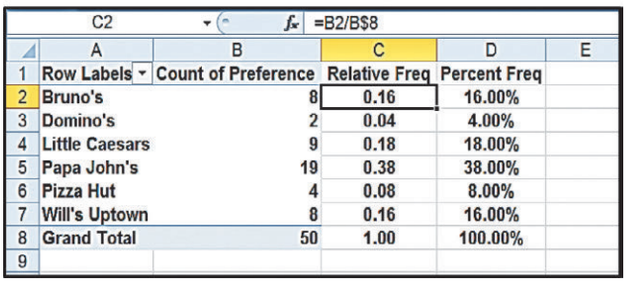
  The PivotChart (bar chart) will also be displayed in a graphics window in the new worksheet.

  Note: You may need to close the PivotTable Field List task pane in order to see the bar chart.



- As demonstrated in Appendix 1.1, move the bar chart to a new worksheet before editing.
- In the new worksheet, the chart can be edited by selecting the **Layout** tab. By clicking on the Labels, Axes, Background, Analysis, and Properties groups, many of the chart characteristics can be edited, data labels (the numbers above the bars that give the bar heights) can be inserted, and so forth. Alternatively, the chart can be edited by right-clicking on various portions of the chart and by using the pop-up menus that are displayed.
- To calculate relative frequencies and percent frequencies of the pizza preferences as in Table 2.3 on page 36, enter the cell formula =B2/B$8 into cell C2 and copy this cell formula down through all of the rows in the PivotTable (that is, through cell C8) to obtain a relative frequency for each row and the total relative frequency of 1.00. Copy the cells containing the relative frequencies into cells D2 through D8, select them, right-click on them, and format the cells to represent percentages to the decimal place accuracy you desire.

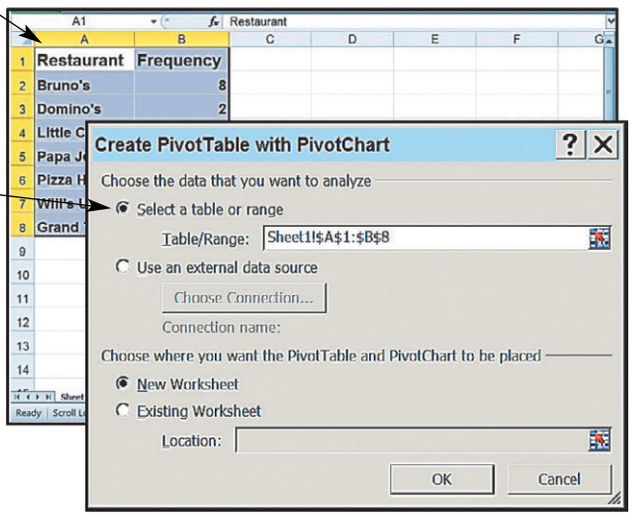




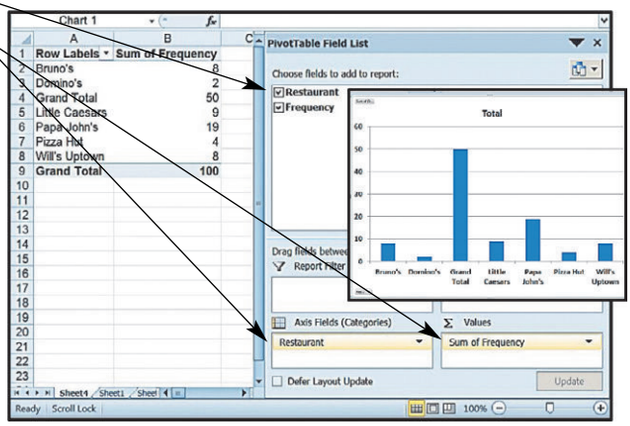
	A	B	C	D	E
1	Row Labels	Count of Preference	Relative Freq	Percent Freq	
2	Bruno's	8	0.16	16.00%	
3	Domino's	2	0.04	4.00%	
4	Little Caesars	9	0.18	18.00%	
5	Papa John's	19	0.38	38.00%	
6	Pizza Hut	4	0.08	8.00%	
7	Will's Uptown	8	0.16	16.00%	
8	Grand Total	50	1.00	100.00%	
9					

Construct a frequency bar chart of pizza preferences from a frequency distribution (data file: PizzaFreq.xlsx):

- Enter the frequency distribution of pizza preferences in Table 2.2 on page 36 as shown in the screen with the various restaurant identifiers in column A (with label Restaurant) and with the corresponding frequencies in column B (with label Frequency).
- Select the entire data set using the mouse.
- Select **Insert : PivotTable : PivotChart**
- In the Create PivotTable with PivotChart dialog box, click "Select a table or range."
- Enter the entire range of the data into the Table/Range window. Here we have entered the range of the frequency distribution—that is, A1 : B8. Again, the easiest way to do this is to click in the Table/Range window and then to select the range of the data with the mouse.
- Select "New Worksheet" to have the PivotTable and PivotChart output displayed in a new worksheet and click OK in the Create PivotTable with PivotChart dialog box.



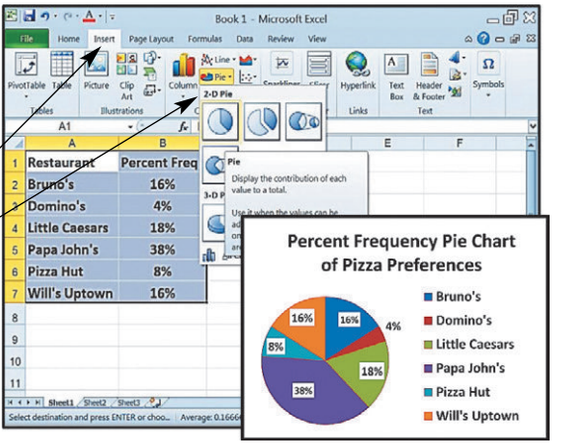
- Place checkmarks in the checkboxes to the left of the fieldnames Restaurant and Frequency. When this is done, the fieldname Restaurant will be placed in the Axis Fields area, "Sum of Frequency" will be placed in the Σ Values area, the Pivot Table will be placed in the new worksheet, and the Pivot Chart (bar chart) will be constructed. You may need to close the PivotTable Field List pane to see the chart. The chart can be moved to a new sheet and edited as previously described.
- **A bar chart can also be constructed without using a pivot table.** To do this, select the entire frequency distribution with the mouse. Then, select Insert: Bar: **All Chart Types**. In the Insert Chart dialog box, select Column from the chart type list on the right, select Clustered Column from the gallery of charts on the right, and click OK.



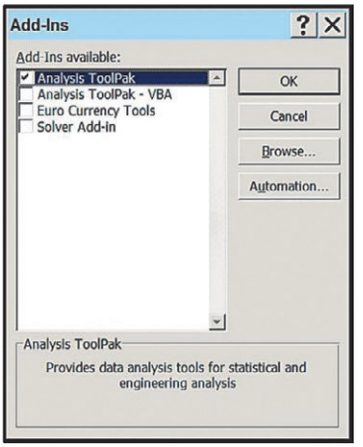
- This method can also be used to construct bar charts of relative frequency and percent frequency distributions. Simply enter the relative or percent distribution into the worksheet, select the entire distribution with the mouse and make the same selections as in the preceeding bullet.

Construct a percentage pie chart of pizza preferences as in Figure 2.3 on page 38 (data file: PizzaPercents.xlsx):

- Enter the percent frequency distribution of pizza preferences in Table 2.3 on page 36 as shown in the screen with the various restaurant identifiers in column A (with label Restaurant) and with the corresponding percent frequencies in column B (with label Percent Freq).
- Select the entire data set using the mouse.
- Select **Insert : Pie : 2-Pie : Pie**
- The pie chart is edited in the same way a bar chart is edited—see the instructions above related to editing bar charts.

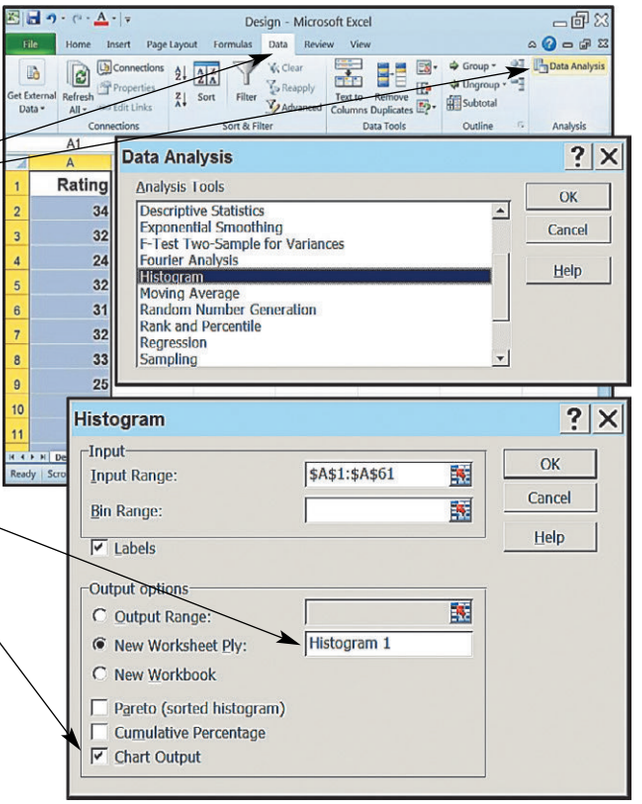


Constructing frequency distributions and histograms using Excel's Analysis ToolPak: The Analysis ToolPak is an Excel add-in that is used for a variety of statistical analyses—including construction of frequency distributions and histograms from raw (that is, un-summarized) data. The ToolPak is available when Microsoft Office or Excel is installed. However, in order to use it, the ToolPak must first be loaded. To see if the Analysis ToolPak has been loaded on your computer, click the **Microsoft File Button,** click **Options,** and finally click **Add-Ins.** If the ToolPak has been loaded on your machine, it will appear in the list of **Active Application Add-ins.** If Analysis ToolPak does not appear in this list, select **Excel Add-ins** in the **Manage** box and click **Go.** In the **Add-ins** box, place a checkmark in the **Analysis ToolPak** checkbox, and then click **OK**. Note that, if the Analysis ToolPak is not listed in the Add-Ins available box, click **Browse** to attempt to find it. If you get prompted that the Analysis ToolPak is not currently installed on your computer, click **Yes** to install it. In some cases, you might need to use your original MS Office or Excel CD/DVD to install and load the Analysis ToolPak by going through the setup process.

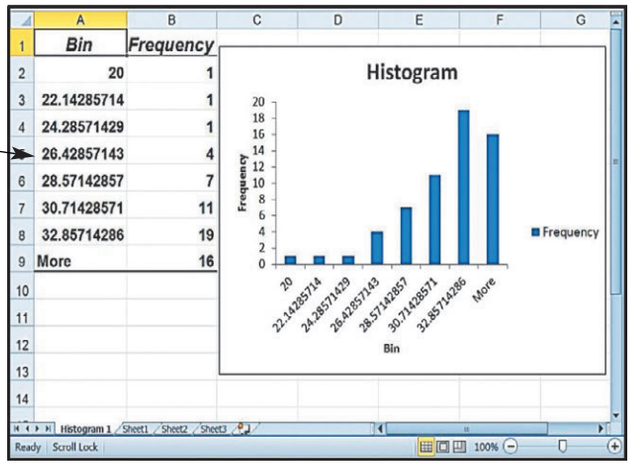


Constructing a frequency histogram of the bottle design ratings as in Figure 2.11 on page 47 (data file: Design.xlsx):

- Enter the 60 bottle design ratings in Table 1.5 on page 10 into Column A with label Rating in cell A1.
- Select **Data : Data Analysis**
- In the Data Analysis dialog box, select Histogram in the Analysis Tools window and click OK.
- In the Histogram dialog box, click in the Input Range window and select the range of the data A1:A61 into the Input Range window by dragging the mouse from cell A1 through cell A61.
- Place a checkmark in the Labels checkbox.
- Under "Output options," select "New Worksheet Ply."
- Enter a name for the new worksheet in the New Worksheet Ply window—here Histogram 1.
- Place a checkmark in the Chart Output checkbox.
- Click OK in the Histogram dialog box.
- Notice that we are leaving the Bin Range window blank. This will cause Excel to define automatic classes for the frequency distribution and histogram. However, because Excel's automatic classes are often not appropriate, we will revise these automatic classes as follows.



- The frequency distribution will be displayed in the new worksheet and the histogram will be displayed in a graphics window.



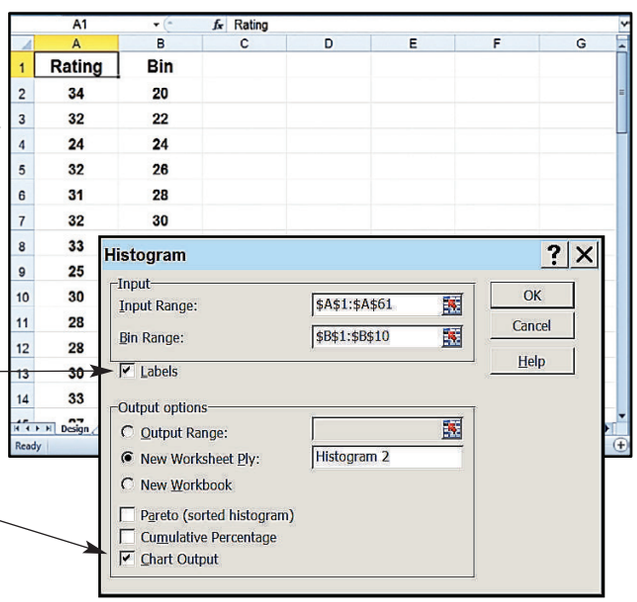
Notice that Excel defines what it calls bins when constructing the histogram. The bins define the automatic classes for the histogram. The bins that are automatically defined by Excel are often cumbersome—the bins in this example are certainly inconvenient for display purposes! Although one might be tempted to simply round the bin values, we have found that the rounded bin values can produce an unacceptable histogram with unequal class lengths (whether this happens depends on the particular bin values in a given situation).

To obtain more acceptable results, **we suggest that new bin values be defined that are roughly based on the automatic bin values.** We can do this as follows. First, we note that the smallest bin value is 20 and that this bin value is expressed using the same decimal place accuracy as the original data (recall that the bottle design ratings are all whole numbers). Remembering that Excel obtains a cell frequency by counting the number of measurements that are less than or equal to the upper class boundary and greater than the lower class boundary, the first class contains bottle design ratings less than or equal to 20. Based on the authors' experience, the first automatic bin value given by Excel is expressed to the same decimal place accuracy as the data being analyzed. However, if the smallest bin value were to be expressed using more decimal places than the original data, then **we suggest rounding it down to the decimal place accuracy of the original data being analyzed.** Frankly, the authors are not sure that this would ever need to be done—it was not necessary in any of the examples we have tried. Next, find the class length of the Excel automatic classes and round it to a convenient value. For the bottle design ratings, using the first and second bin values in the screen, the class length is 22.14285714 – 20 which equals 2.14285714. To obtain more convenient classes, we will round this value to 2. Starting at the first automatic bin value of 20, we now construct classes having length equal to 2. This gives us new bin values of 20, 22, 24, 26, and so on. We suggest continuing to define new bin values until a class containing the largest measurement in the data is found. Here, the largest bottle design rating is 35 (see Table 1.5 on page 10). Therefore, the last bin value is 36, which says that the last class will contain ratings greater than 34 and less than or equal to 36—that is, the ratings 35 and 36.

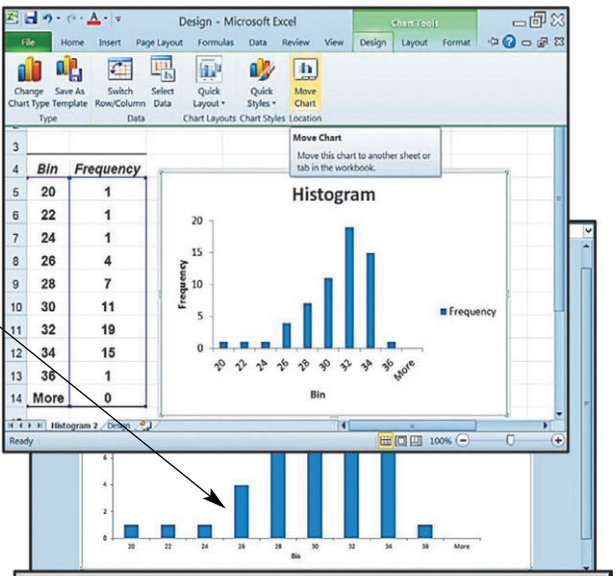
We suggest constructing classes in this way unless one or more measurements are unusually large compared to the rest of the data—we might call these unusually large measurements **outliers.** We will discuss outliers more thoroughly in Chapter 3 (and in later chapters). For now, if we (subjectively) believe that one or more outliers exist, we suggest placing these measurements in the "more" class and placing a histogram bar over this class having the same class length as the other bars. In such a situation, we must recognize that the Excel histogram will not be technically correct because **the area of the bar (or rectangle) above the "more" class will not necessarily equal the relative proportion of measurements in the class.** Nevertheless—given the way Excel constructs histogram classes—the approach we suggest seems reasonable. In the bottle design situation, the largest rating of 35 is not unusually large and, therefore, the "more" class will not contain any measurements.

To construct the revised histogram:

- Open a new worksheet, copy the bottle design ratings into column A and enter the new bin values into column B (with label Bin) as shown.
- Select **Data : Data Analysis : Histogram**
- Click OK in the Data Analysis dialog box.
- In the Histogram dialog box, select the range of the ratings data A1:A61 into the Input Range window.
- Click in the Bin Range window and enter the range of the bin values B1:B10.
- Place a checkmark in the Labels checkbox.
- Under "Output options," select "New Worksheet Ply" and enter a name for the new worksheet—here Histogram 2.
- Place a checkmark in the Chart Output checkbox.
- Click OK in the Histogram dialog box.

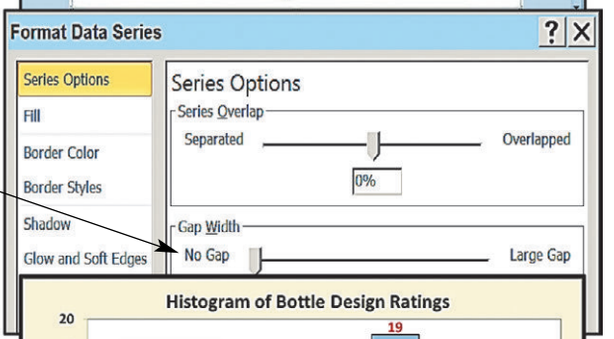


- The revised frequency distribution will be displayed in the new worksheet and the histogram will be displayed in a graphics window.
- Click in the graphics window and (as demonstrated in Appendix 1.1) move the histogram to a new worksheet for editing.
- The histogram will be displayed in the new chart sheet in a much larger format that makes it easier to carry out editing.

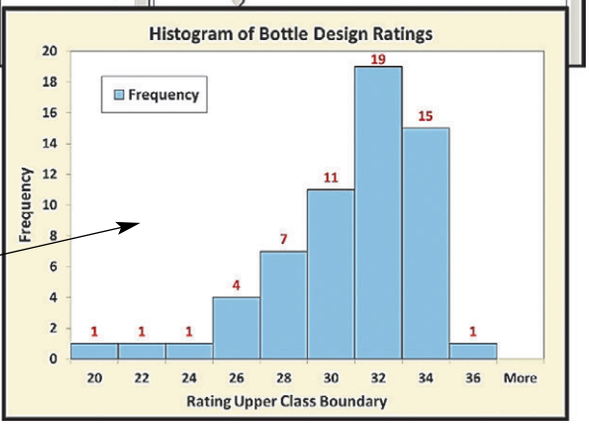


To remove the gaps between the histogram bars:

- Right click on one of the histogram bars and select Format Data Series from the pop-up window.
- Set the gap width to zero by moving the gap width slider to "No Gap" and click "Close" in the Format Data Series dialog box.
- By selecting the Chart Tools Layout tab, the histogram can be edited in many ways. This can also be done by right clicking on various portions of the histogram and by making desired pop-up menu selections.
- To obtain **data labels** (the numbers on the tops of the bars that indicate the bar heights), right click on one of the histogram bars and select "Add data labels" from the pop-up menu.

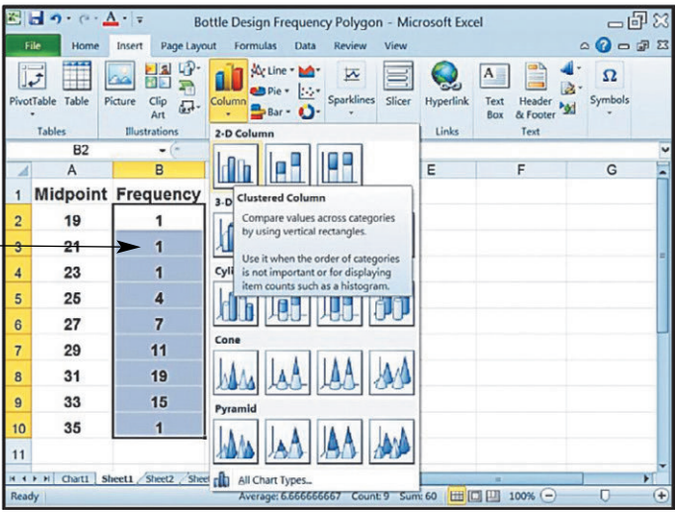


After final editing, the histogram might look like the one illustrated in Figure 2.11 on page 47.

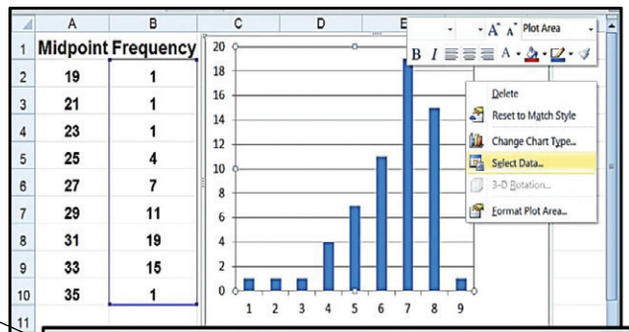


Constructing a frequency histogram of bottle design ratings from summarized data:

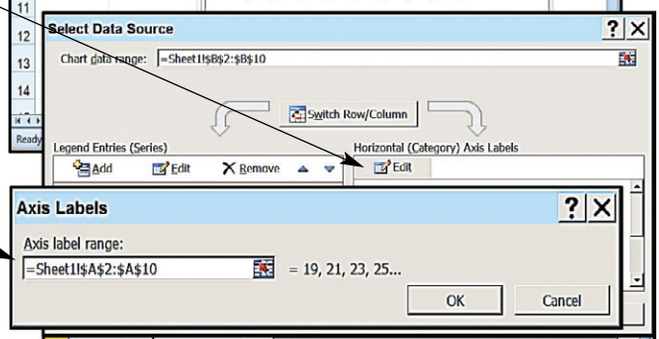
- Enter the **midpoints** of the frequency distribution classes into column A with label Midpoint and enter the class frequencies into column B with label Frequency.
- Use the mouse to select the cell range that contains the frequencies (here, cells B2 through B10).
- Select **Insert : Column : 2-D Column (Clustered Column)**



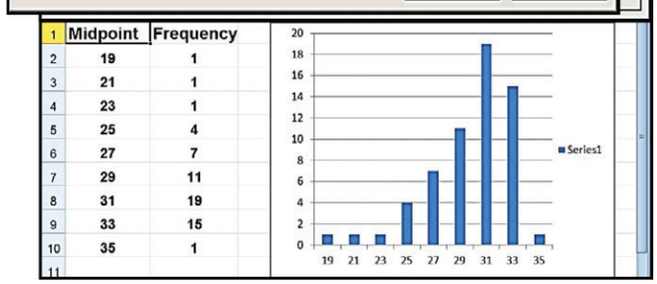
- Right-click on the chart that is displayed and click on Select Data in the pop-up menu.



- In the Select Data Source dialog box, click on the Edit button in the "Horizontal (Category) Axis Labels" window.



- In the Axis Labels dialog box, use the mouse to enter the cell range that contains the midpoints (here, A2:A10) into the "Axis label range" window.
- Click OK in the Axis Labels dialog box.
- Click OK in the Select Data Source dialog box.
- Move the chart that appears to a chart sheet, remove the gaps between the bars as previously shown, and edit the chart as desired.

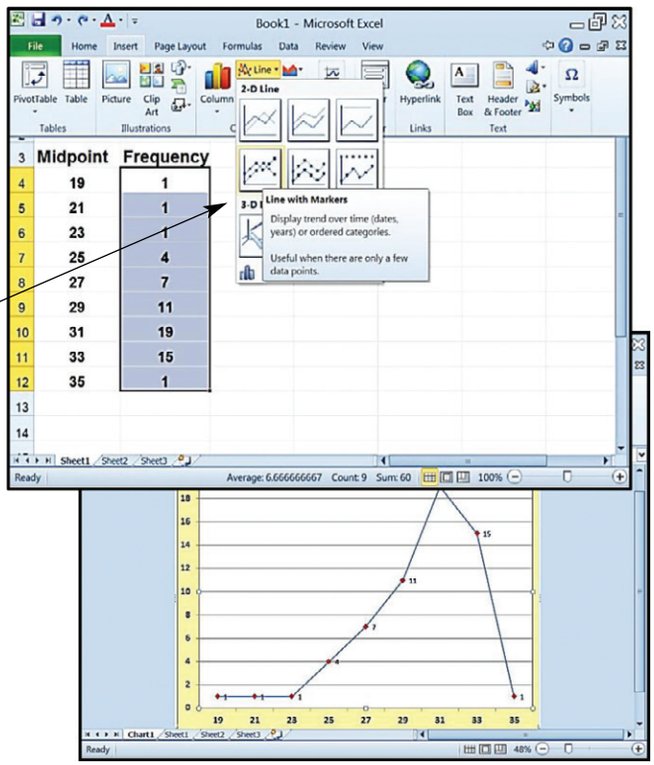


Relative frequency or percent frequency histograms would be constructed in the same way with the class midpoints in column A of the Excel spreadsheet and with the relative or percent frequencies in column B.

We now show how to construct a **frequency polygon from summarized data.**

Note that, if the data are **not summarized,** first use the Histogram option in the Analysis ToolPak to develop a summarized frequency distribution.

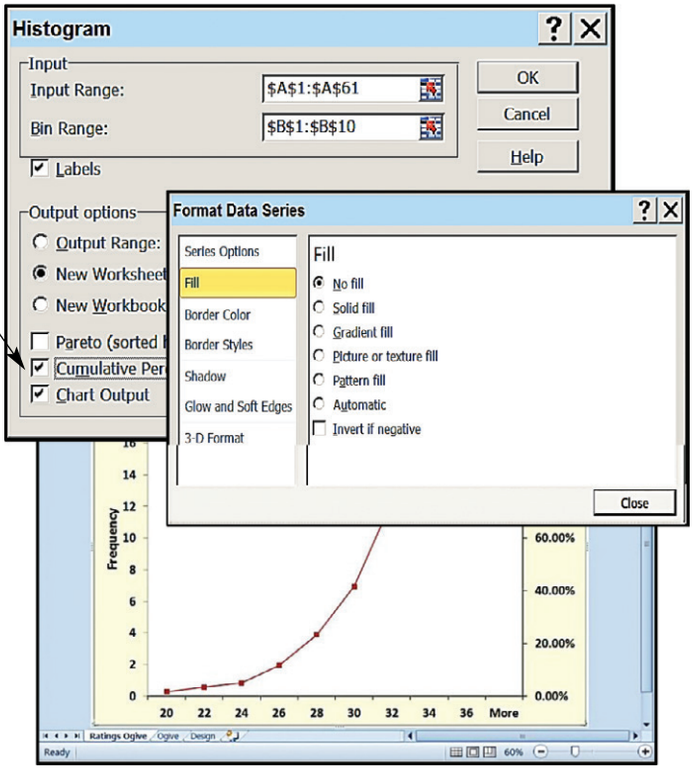
- Enter the class midpoints and class frequencies as shown previously for summarized data.
- Use the mouse to select the cell range that contains the frequencies.
- Select **Insert : Line : Line with markers**
- Right-click on the chart that is displayed and click on Select Data in the pop-up menu.
- In the Select Data Source dialog box, click on the Edit button in the "Horizontal (Category) Axis Labels" window.
- In the Axis Labels dialog box, use the mouse to enter the cell range that contains the midpoints into the "Axis label range" window.
- Click OK in the Axis Labels dialog box.
- Click OK in the Select Data Source dialog box.
- Move the chart that appears to a chart sheet, and edit the chart as desired.



To construct a percent frequency ogive for the bottle design rating distribution (data file: Design.xlsx):

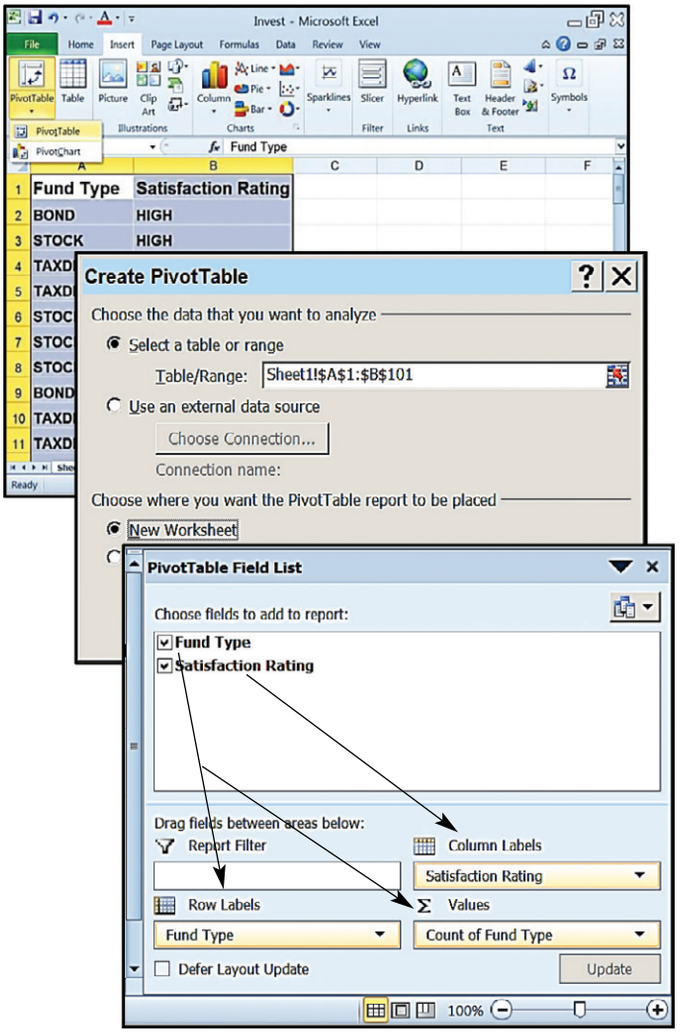
Follow the instructions for constructing a histogram by using the Analysis ToolPak with changes as follows:

- In the Histogram dialog box, place a checkmark in the Cumulative Percentage checkbox.
- After moving the histogram to a chart sheet, right-click on any histogram bar.
- Select "Format Data Series" from the pop-up menu.
- In the "Format Data Series" dialog box, (1) select Fill from the list of "Series Options" and select "No fill" from the list of Fill options; (2) select Border Color from the list of "Series Options" and select "No line" from the list of Border Color options; (3) Click Close.
- Click on the chart to remove the histogram bars.

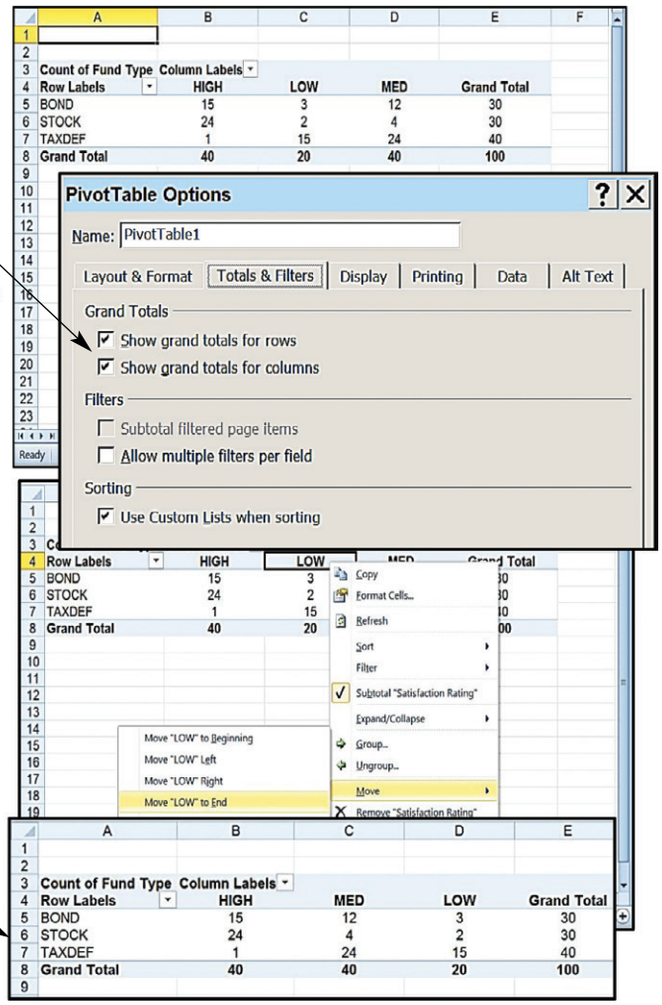


Construct a contingency table of fund type versus level of client satisfaction as in Table 2.17 on page 62 (data file: Invest.xlsx):

- Enter the customer satisfaction data in Table 2.16 on page 62—fund types in column A with label "Fund Type" and satisfaction ratings in column B with label "Satisfaction Rating."
- Select **Insert : PivotTable : PivotTable**
- In the Create PivotTable dialog box, click "Select a table or range."
- By dragging with the mouse, enter the range of the data to be analyzed into the Table/Range window. Here we have entered the range of the client satisfaction data A1:B101.
- Select the New Worksheet option to place the PivotTable in a new worksheet.
- Click OK in the Create PivotTable dialog box.
- In the PivotTable Field List task pane, drag the label "Fund Type" and drop it into the Row Labels area.
- Also drag the label "Fund Type" and drop it into the Σ Values area. When this is done, the label will automatically change to "Count of Fund Type."
- Drag the label "Satisfaction Rating" into the Column Labels area.

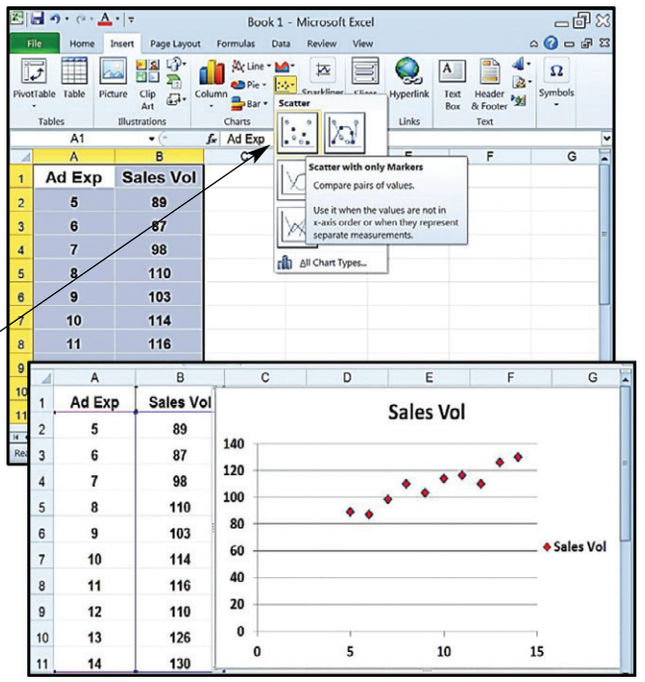


- The PivotTable will be created and placed in a new worksheet.
- Now right-click inside the PivotTable and select PivotTable Options from the pop-up menu.
- In the PivotTable Options dialog box, select the Totals & Filters tab and make sure that a checkmark has been placed in each of the "Show grand totals for rows" and the "Show grand totals for columns" checkboxes.
- Select the Layout & Format tab, place a checkmark in the "For empty cells show" checkbox and enter 0 (the number zero) into its corresponding window. (For the customer satisfaction data, none of the cell frequencies equal zero, but, in general, this setting should be made to prevent empty cells from being left blank in the contingency table.)
- To change the order of the column labels from the default alphabetical ordering (High, Low, Medium) to the more logical ordering of High, Medium, Low, right-click on LOW, select Move from the pop-up menu, and select "Move LOW to End."
- The contingency table is now complete.



Constructing a scatter plot of sales volume versus advertising expenditure as in Figure 2.24 on page 67 (data file: SalesPlot.xlsx):

- Enter the advertising and sales data in Table 2.20 on page 67 into columns A and B—advertising expenditures in column A with label "Ad Exp" and sales values in column B with label "Sales Vol." **Note: The variable to be graphed on the horizontal axis must be in the first column** (that is, the left-most column) and **the variable to be graphed on the vertical axis must be in the second column** (that is, the rightmost column).
- Select the entire range of data to be graphed.
- Select **Insert : Scatter : Scatter with only Markers**
- The scatter plot will be displayed in a graphics window. Move the plot to a chart sheet and edit appropriately.

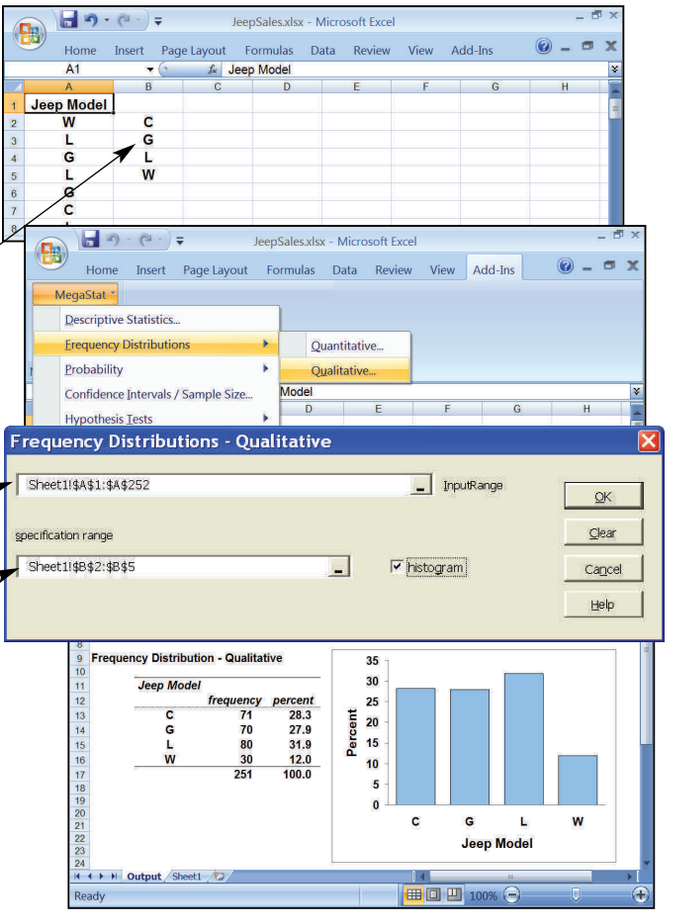


Appendix 2.2 ■ Tabular and Graphical Methods Using MegaStat

The instructions in this section begin by describing the entry of data into an Excel worksheet. Alternatively, the data may be downloaded from this book's website. The appropriate data file name is given at the top of each instruction block. Please refer to Appendix 1.1 for further information about entering data, saving data, and printing results in Excel. Please refer to Appendix 1.2 for more information about MegaStat basics.

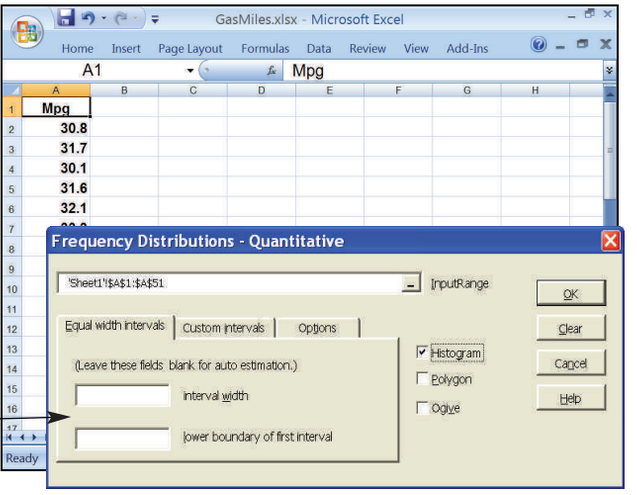
Construct a frequency distribution and bar chart of Jeep sales in Table 2.23 on page 74 (data file: JeepSales.xlsx):

- Enter the Jeep sales data in Table 2.23 on page 74 (C = Commander; G = Grand Cherokee; L = Liberty; W = Wrangler) into column A with label Jeep Model in cell A1.
- Enter the categories for the qualitative variable (C, G, L, W) into the worksheet. Here we have placed them in cells B2 through B5—the location is arbitrary.
- Select **Add-Ins : MegaStat : Frequency Distributions : Qualitative**
- In the "Frequency Distributions—Qualitative" dialog box, use the AutoExpand feature to enter the range A1:A252 of the Jeep sales data into the Input Range window.
- Enter the cell range B2:B5 of the categories (C, G, L, W) into the "specification range" window.
- Place a checkmark in the "histogram" checkbox to obtain a bar chart.
- Click OK in the "Frequency Distributions—Qualitative" dialog box.
- The frequency distribution and bar chart will be placed in a new output sheet.
- The output can be edited in the output sheet. Alternatively, the bar chart can be moved to a chart sheet (see Appendix 1.1) for more convenient editing.



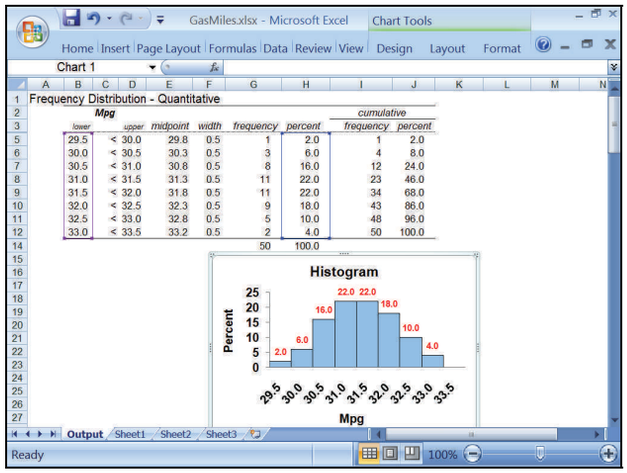
Construct a frequency distribution and percent frequency histogram of the gas mileages as in Figure 2.9 on page 46 (data file: GasMiles.xlsx):

- Enter the gasoline mileage data in Table 1.6 on page 11 into column A with the label Mpg in cell A1 and with the 50 gas mileages in cells A2 to A51.
- Select **Add-Ins : MegaStat : Frequency Distributions : Quantitative**
- In the "Frequency Distributions—Quantitative" dialog box, use the AutoExpand feature to enter the range A1:A51 of the gas mileages into the Input Range window.
- To obtain **automatic classes** for the histogram, leave the "interval width" and "lower boundary of first interval" windows blank.
- Place a checkmark in the Histogram checkbox.
- Click OK in the "Frequency Distributions—Quantitative" dialog box.



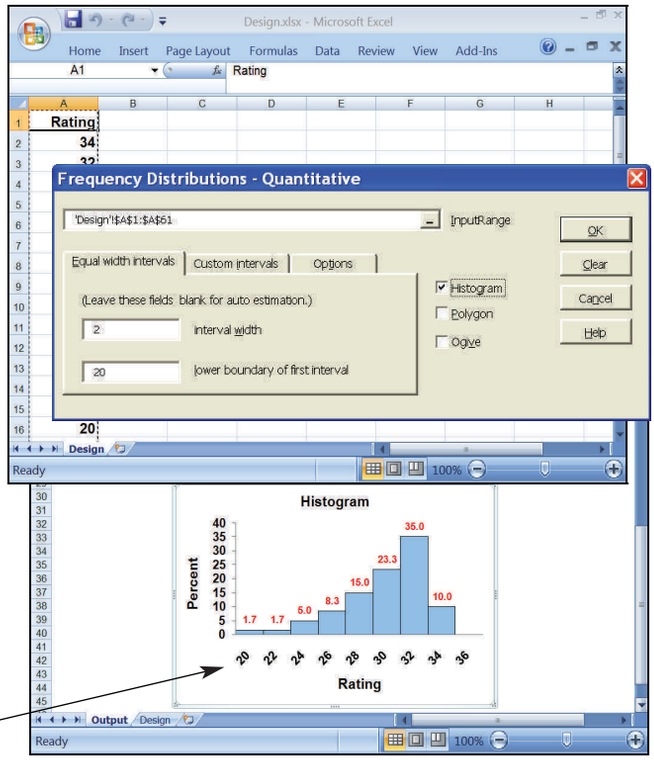
- The frequency distribution and histogram will be placed in a new output worksheet.
- The chart can be edited in the Output worksheet or you can move the chart to a chart sheet for editing.
- To obtain **data labels** (the numbers on the tops of the bars that indicate the bar heights), right click on one of the histogram bars and select "Add data labels" from the pop-up menu.

To construct a frequency polygon and a percent frequency ogive, simply place checkmarks in the Polygon and Ogive checkboxes in the "Frequency Distributions—Quantitative" dialog box.



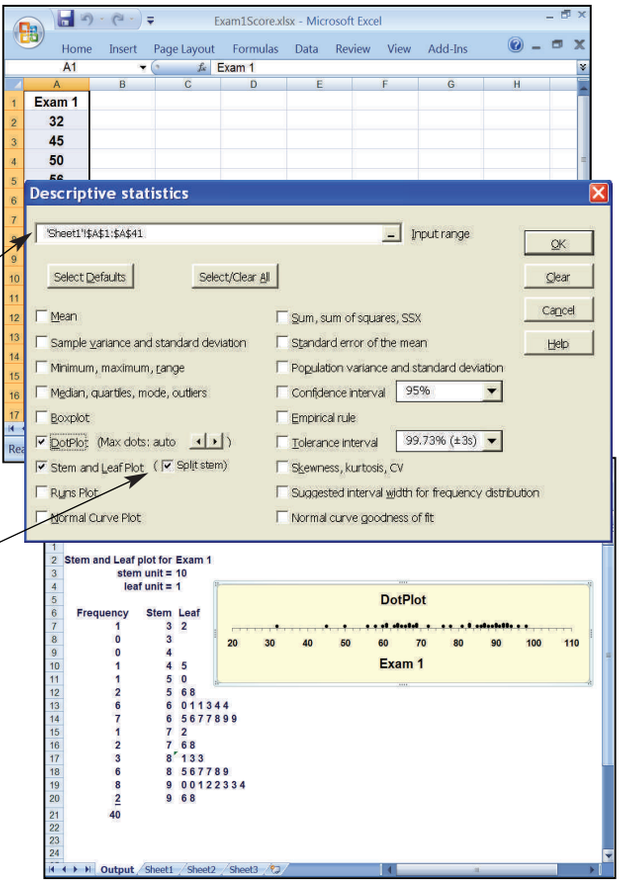
Construct a percent frequency histogram of the bottle design ratings similar to Figure 2.11 on page 47 with **user specified classes** (data file: Design.xlsx):

- Enter the 60 bottle design ratings in Table 1.5 on page 10 into Column A with label Rating in cell A1.
- Select **Add-Ins : MegaStat : Frequency Distributions : Quantitative**
- In the "Frequency Distributions—Quantitative" dialog box, use the AutoExpand feature to enter the input range A1:A61 of the bottle design ratings into the Input Range window.
- Enter the class width (in this case equal to 2) into the "interval width" window.
- Enter the lower boundary of the first—that is, leftmost—class (or interval) of the histogram (in this case equal to 20) into the "lower boundary of first interval" window.
- Make sure that the Histogram checkbox is checked.
- Click OK in the "Frequency Distributions—Quantitative" dialog box.
- We obtain a histogram with class boundaries 20, 22, 24, 26, 28, 30, 32, 34, and 36. Note that the appearance of this histogram is not exactly the same as that of the Excel histogram in Figure 2.11 on page 47 because MegaStat and Excel count frequencies differently. While MegaStat counts frequencies as we have described in Example 2.2, recall that Excel counts the number of measurements that are greater than the lower boundary of a class and less than or equal to the upper boundary of the class.
- The histogram can be moved to a chart sheet for editing purposes.



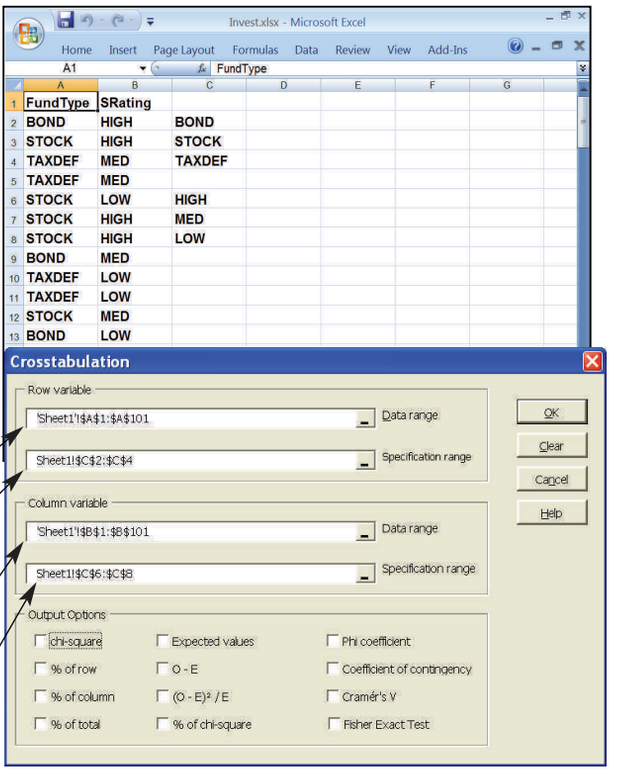
Construct a dot plot (as in Figure 2.18 on page 55) **and a stem-and-leaf display** (as in Figure 2.20 on page 59) of the scores on the first statistics exam as discussed in Example 2.3 on page 48 (data file: FirstExam.xlsx):

- Enter the 40 scores for exam 1 in Table 2.8 on page 48 into column A with label "Exam 1" in cell A1.
- Select **Add-Ins : MegaStat : Descriptive Statistics.**
- In the "Descriptive Statistics" dialog box, use the AutoExpand feature to enter the range A1:A41 of the exam scores into the "Input range" window.
- Place a checkmark in the DotPlot checkbox to obtain a dot plot.
- Place a checkmark in the "Stem and Leaf Plot" checkbox to obtain a stem-and-leaf display.
- Place a checkmark in the "Split Stem" checkbox. (In general, whether or not this should be done depends on how you want the output to appear. You may wish to construct two plots—one with the Split Stem option and one without—and then choose the output you like best.) In the exam score situation, the Split Stem option is needed to obtain a display that looks like the one in Figure 2.20.
- Click OK in the "Descriptive Statistics" dialog box.
- The dot plot and stem-and-leaf display will be placed in an output sheet. Here, the stem-and-leaf display we have obtained for exam 1 is the "mirror image" of the display shown in Figure 2.20 (because we have constructed a single display for exam 1, while Figure 2.20 shows back-to-back displays for both exams 1 and 2).
- The dot plot can be moved to a chart sheet for editing.



Construct a contingency table of fund type versus level of client satisfaction as in Table 2.17 on page 62 (data file: Invest.xlsx):

- Enter the customer satisfaction data in Table 2.16 on page 62—fund types in column A with label FundType and satisfaction ratings in column B with label SRating.
- Enter the three labels (BOND; STOCK; TAXDEF) for the qualitative variable FundType into cells C2, C3, and C4 as shown in the screen.
- Enter the three labels (HIGH; MED; LOW) for the qualitative variable SRating into cells C6, C7, and C8 as shown in the screen.
- Select **Add-Ins : MegaStat : Chi-Square/CrossTab : Crosstabulation**
- In the Crosstabulation dialog box, use the AutoExpand feature to enter the range A1:A101 of the row variable FundType into the "Row variable Data range" window.
- Enter the range C2:C4 of the labels of the qualitative variable FundType into the "Row variable Specification range window."
- Use the AutoExpand feature to enter the range B1:B101 of the column variable SRating into the "Column variable Data range" window.
- Enter the range C6:C8 of the labels of the qualitative variable SRating into the "Column variable Specification range window."



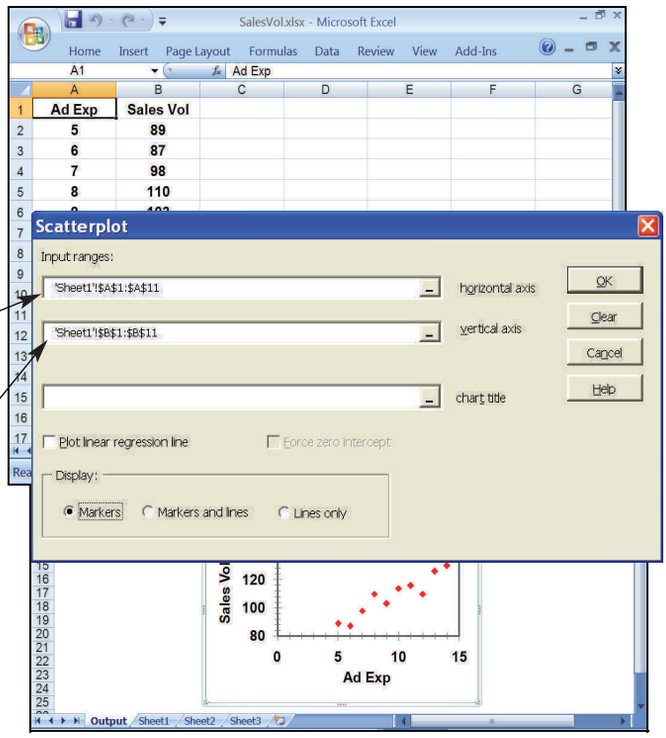
- Uncheck the "chi-square" checkbox.
- Click OK in the Crosstabulation dialog box.
- The contingency table will be displayed in an Output worksheet.
- **Row percentages and column percentages** can be obtained by simply placing checkmarks in the "% of row" and "% of column" checkboxes.

Crosstabulation

		SRating			
		HIGH	MED	LOW	Total
FundType	BOND	15	12	3	30
	STOCK	24	4	2	30
	TAXDEF	1	24	15	40
	Total	40	40	20	100

Construct a scatter plot of sales volume versus advertising expenditure as in Figure 2.24 on page 67 (data file: SalesPlot.xlsx):

- Enter the advertising and sales data in Table 2.20 on page 67 into columns A and B—advertising expenditures in column A with label "Ad Exp" and sales values in column B with label "Sales Vol."
- Select **Add-Ins : MegaStat : Correlation/Regression : Scatterplot**
- In the Scatterplot dialog box, use the AutoExpand feature to enter the range A1:A11 of the advertising expenditures into the "horizontal axis" window.
- Use the AutoExpand feature to enter the range B1:B11 of the sales volumes into the "vertical axis" window.
- Uncheck the "Plot linear regression line" checkbox.
- Under Display options, select Markers.
- Click OK in the Scatterplot dialog box.
- The scatterplot is displayed in an Output worksheet and can be moved to a chart sheet for editing.



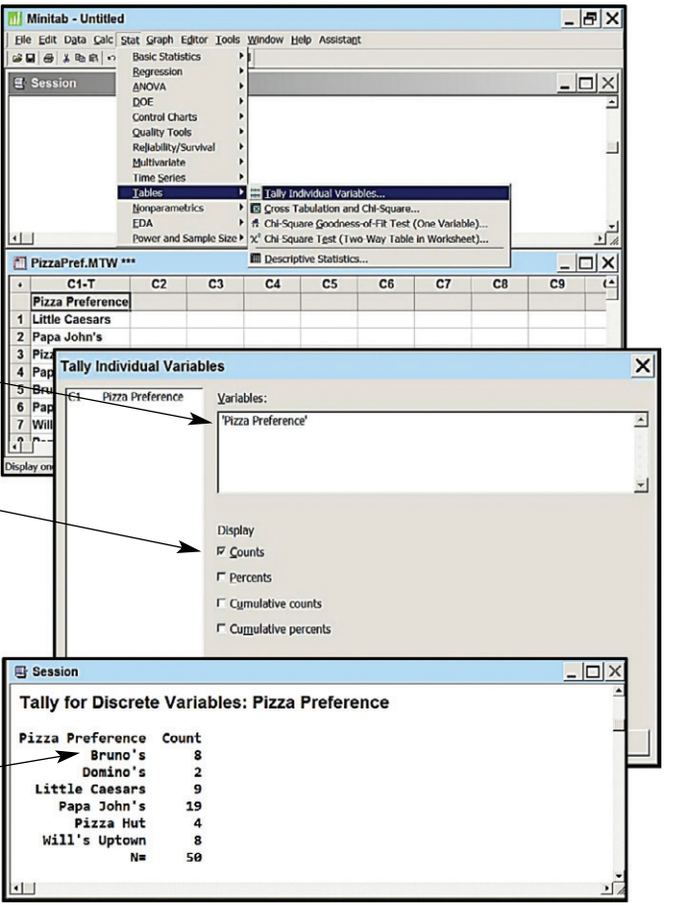
Appendix 2.3 ■ Tabular and Graphical Methods Using MINITAB

The instructions in this section begin by describing the entry of data into the MINITAB data window. Alternatively, the data may be downloaded from this book's website. The appropriate data file name is given at the top of each instruction block. Please refer to Appendix 1.3 for further information about entering data, saving data, and printing results when using MINITAB.

Construct a frequency distribution of pizza preferences as in Table 2.2 on page 36 (data file: PizzaPref.MTW):

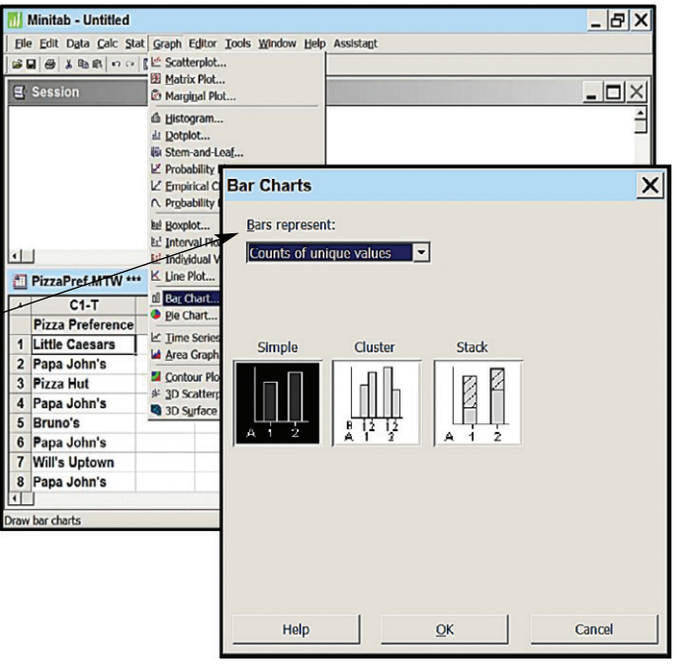
- Enter the pizza preference data in Table 2.1 on page 36 into column C1 with label (variable name) Pizza Preference.
- Select **Stat : Tables : Tally Individual Variables**
- In the Tally Individual Variables dialog box, enter the variable name 'Pizza Preference' into the Variables window. Because this variable name consists of more than one word, we must enclose the name in single quotes—this defines both the words Pizza and Preference to be parts of the same variable name.
- Place a checkmark in the Display "Counts" checkbox to obtain frequencies.

  We would check "Percents" to obtain percent frequencies, "Cumulative counts" to obtain cumulative frequencies, and "Cumulative percents" to obtain cumulative percent frequencies.
- Click OK in the Tally Individual Variables dialog box.
- The frequency distribution is displayed in the Session window.

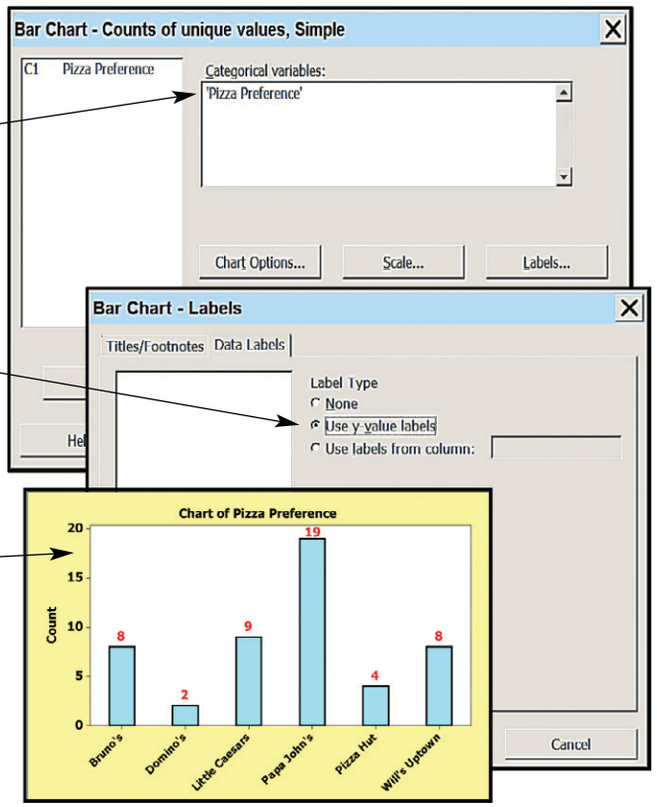


Construct a bar chart of the pizza preference distribution from the raw preference data similar to the bar chart in Figure 2.1 on page 37 (data file: PizzaPref.MTW):

- Enter the pizza preference data in Table 2.1 on page 36 into column C1 with label (variable name) Pizza Preference.
- Select **Graph : Bar Chart**
- In the Bar Charts dialog box, select "Counts of unique values" from the "Bars represent" pull-down menu.
- Select "Simple" from the gallery of bar chart types (this is the default selection, which is indicated by the reverse highlighting in black).
- Click OK in the Bar Charts dialog box.

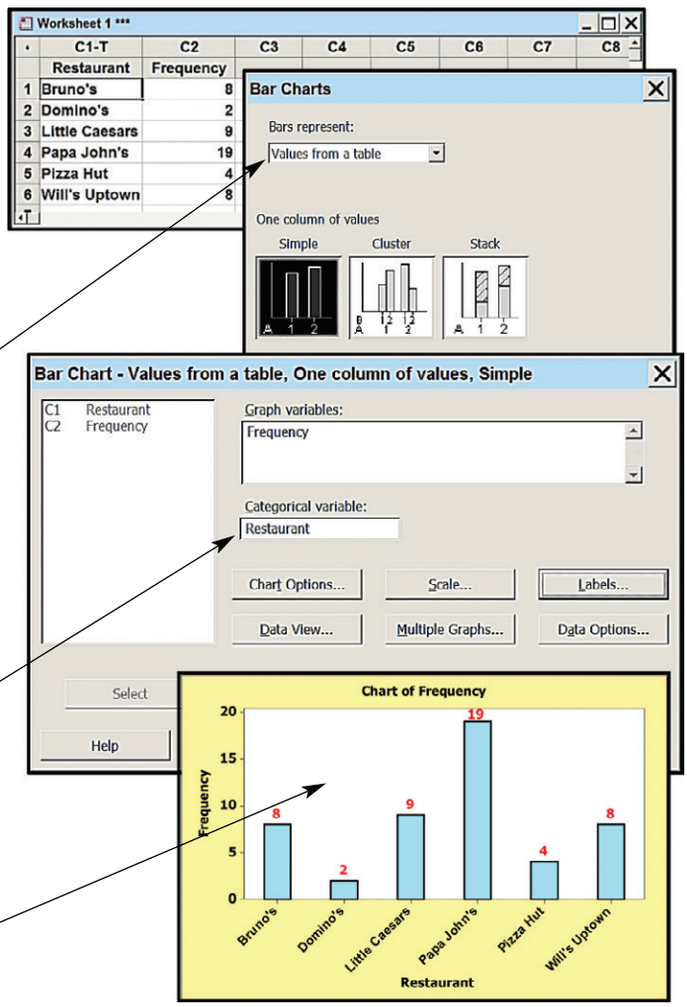


- In the "Bar Chart—Counts of unique values, Simple" dialog box, enter the variable name 'Pizza Preference' into the "Categorical variables" window. Be sure to remember the single quotes around the name Pizza Preference.
- To obtain **data labels** (numbers at the tops of the bars that indicate the heights of the bars—in this case, the frequencies), click on the Labels... button.
- In the "Bar Chart—Labels" dialog box, click on the Data Labels tab and select "Use y-value labels". This will produce data labels that are equal to the category frequencies.
- Click OK in the "Bar Chart—Labels" dialog box.
- Click OK in the "Bar Chart—Counts of unique values, Simple" dialog box.
- The bar chart will be displayed in a graphics window. The chart may be edited by right-clicking on various portions of the chart and by using the pop-up menus that appear—see Appendix 1.3 for more details.
- Here we have obtained a frequency bar chart. **To obtain a percent frequency bar chart** (as in Figure 2.2 on page 37) click on the Chart Options... button and select "Show Y as Percent" in the "Bar Chart—Options" dialog box.



Construct a bar chart from the tabular frequency distribution of pizza preferences in Table 2.2 on page 36 (data file: PizzaFreq.MTW):

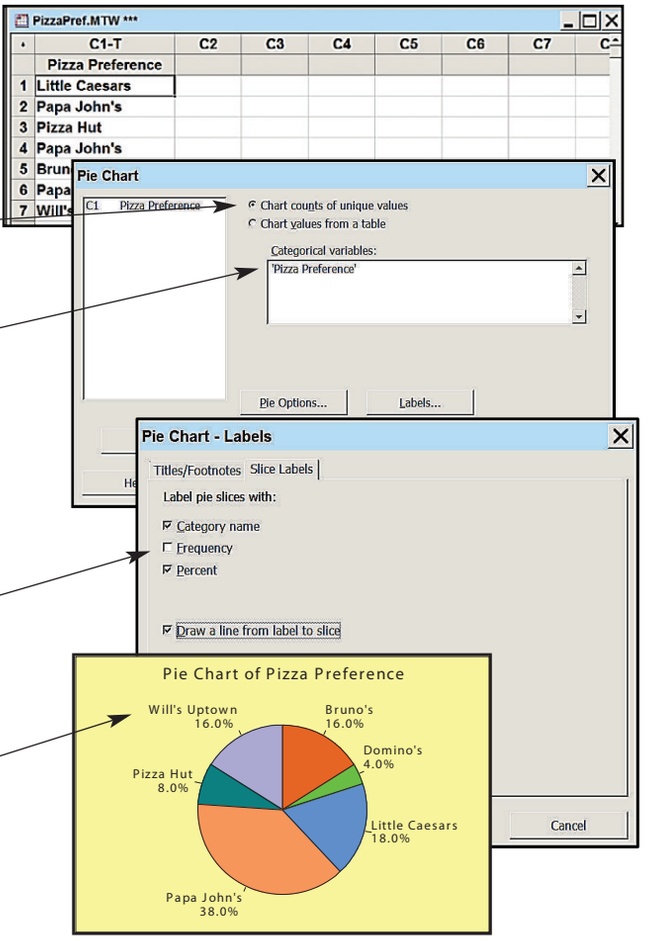
- Enter the tabular distribution of pizza preferences from Table 2.2 as shown in the screen with the pizza restaurants in column C1 (with the variable name Restaurant) and with the associated frequencies in column C2 (with variable name Frequency).
- Select **Graph : Bar Chart**
- In the Bar Charts dialog box, select "Values from a table" in the "Bars represent" pull-down menu.
- Select "One column of values—Simple" from the gallery of bar chart types.
- Click OK in the Bar Charts dialog box.
- In the "Bar Chart—Values from a table, One column of values, Simple" dialog box, enter the variable name Frequency into the "Graph variables" window and then enter the variable name Restaurant into the "Categorical variable" window.
- Click on the Labels... button and select "Use y-value labels" as shown previously.
- Click OK in the "Bar Chart—Labels" dialog box.
- Click OK in the "Bar Chart—Values from a table, One column of values, Simple" dialog box.
- The bar chart will be displayed in a graphics window.



Construct a pie chart of pizza preference percentages similar to that shown in Figure 2.3 on page 38.

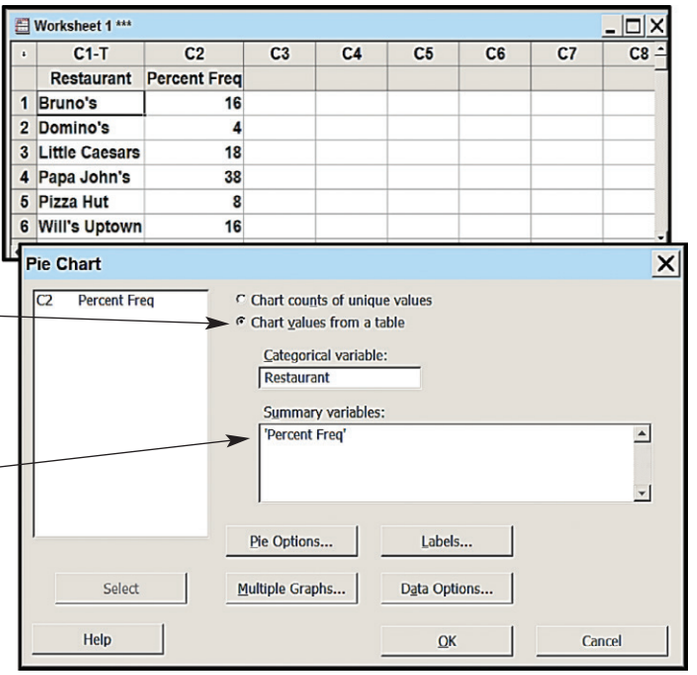
- Enter the pizza preference data in Table 2.1 on page 36 into column C1 with label (variable name) Pizza Preference.
- Select **Graph : Pie Chart**
- In the Pie Chart dialog box, select "Chart counts of unique values."
- Enter the variable name 'Pizza Preference' into the "Categorical variables" window. Be sure to remember the single quotes around the name Pizza Preference.
- In the Pie Chart dialog box, click on the Labels . . . button.
- In the "Pie Chart—Labels" dialog box, click on the Slice Labels tab.
- Place checkmarks in the Category name, Percent, and "Draw a line from label to slice" checkboxes.

  To obtain a frequency pie chart, select Frequency rather than Percent in this dialog box. Or, both Percent and Frequency can be selected.
- Click OK in the "Pie Chart—Labels" dialog box.
- Click OK in the Pie Chart dialog box.
- The pie chart will appear in a graphics window.



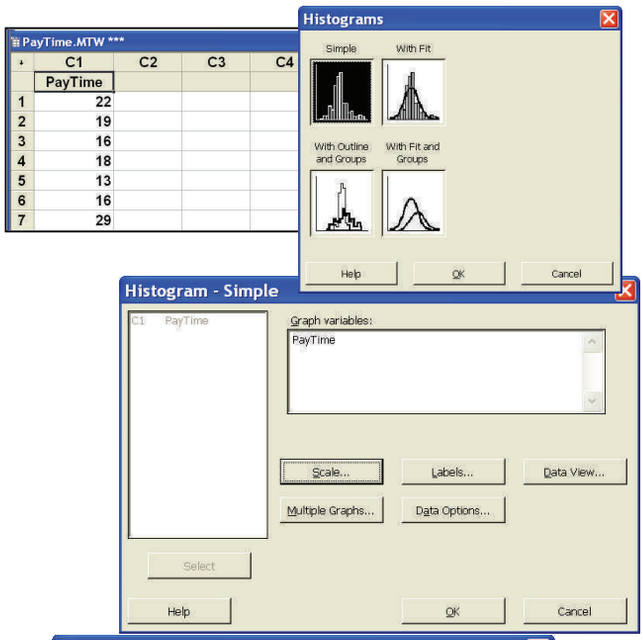
Construct a pie chart from the tabular frequency distribution of pizza preferences in Table 2.3 on page 36 (data file: PizzaPercents.MTW):

- Enter the pizza preference percent frequency distribution from Table 2.3 as shown in the screen with the pizza restaurants in column C1 (with variable name Restaurant) and with the associated percent frequencies in column C2 (with variable name Percent Freq).
- Select **Graph : Pie Chart**
- In the Pie Chart dialog box, select "Chart values from a table."
- Enter the variable name Restaurant into the "Categorical variable" window.
- Enter the variable name 'Percent Freq' into the "Summary variables" window. Be sure to remember the single quotes around the name Percent Freq.
- Continue by following the previously given directions for adding data labels and for generating the pie chart.

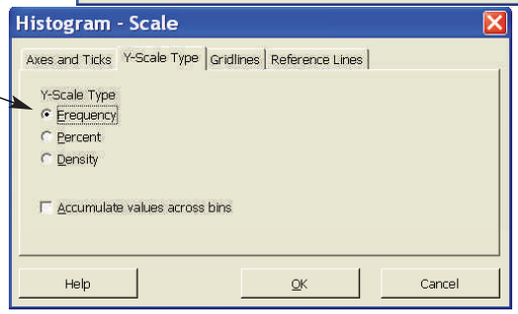


Construct a frequency histogram of the payment times in Figure 2.10 on page 46 (data file: PayTime .MTW):

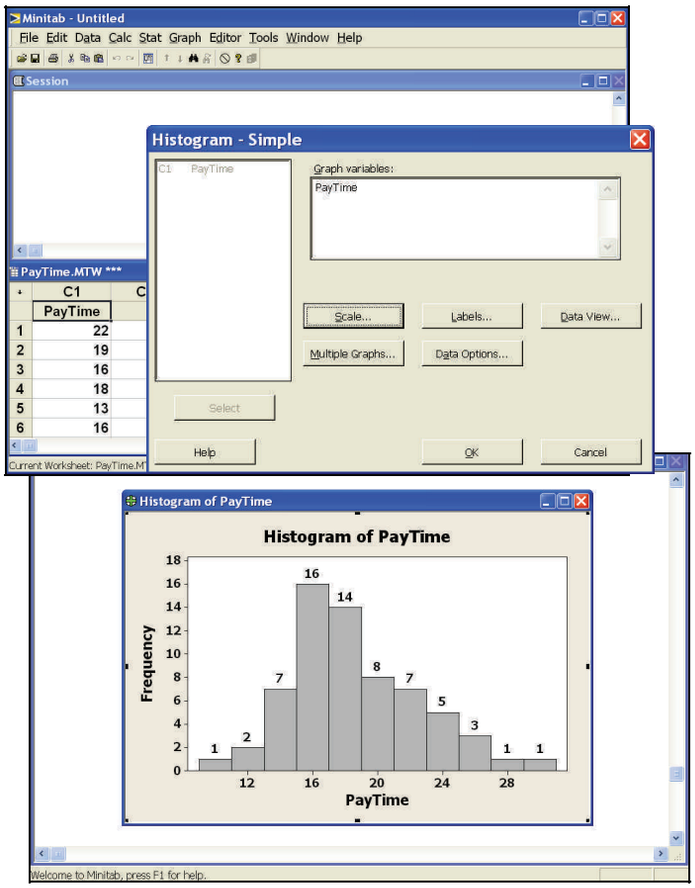
- Enter the payment time data from Table 2.4 on page 42 into column C1 with variable name PayTime.
- Select **Graph : Histogram**
- In the Histograms dialog box, select Simple from the gallery of histogram types and click OK.
- In the "Histogram—Simple" dialog box, enter the variable name PayTime into the Graph Variables window and click on the Scale button.



- In the "Histogram—Scale" dialog box, click on the "Y- Scale Type" tab and select **Frequency** to obtain a frequency histogram. We would select **Percent** to request a percent frequency histogram. Then click OK in the "Histogram—Scale" dialog box.
- **Data labels** are requested in the same way as we have demonstrated for bar charts. Click on the Labels… button in the "Histogram—Simple" dialog box. In the "Histogram—Labels" dialog box, click on the Data Labels tab and select "Use y-value labels." Then click OK in the "Histogram—Labels" dialog box.



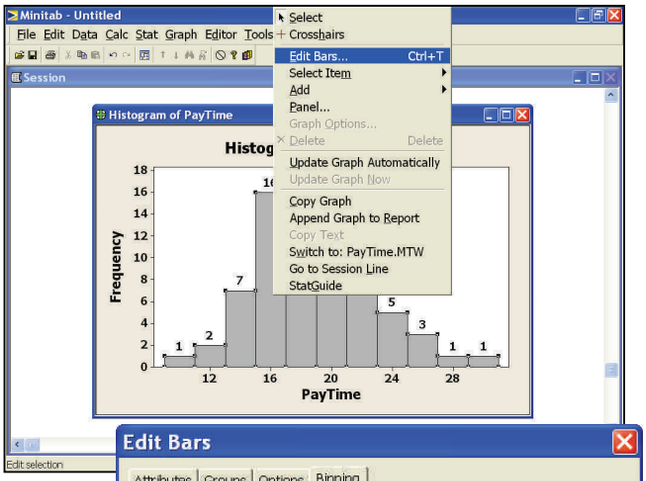
- To create the histogram, click OK in the "Histogram—Simple" dialog box.
- The histogram will appear in a graphics window and can be edited as described in Appendix 1.3.
- The histogram can be selected for printing or can be copied and pasted into a word processing document. (See Appendix 1.3.)



- Notice that MINITAB **automatically** defines classes for the histogram bars, and automatically provides labeled tick marks (here 12, 16, 20, 24 and 28) on the *x*-scale of the histogram. These automatic classes are not the same as those we formed in Example 2.2, summarized in Table 2.7, and illustrated in Figure 2.7 on page 44. However, we can edit the automatically constructed histogram to produce the histogram classes of Figure 2.7. This is sometimes called **"binning."**

To obtain user specified histogram classes—for example, the payment time histogram classes of Figure 2.7 on page 44 (data file: PayTime.MTW):

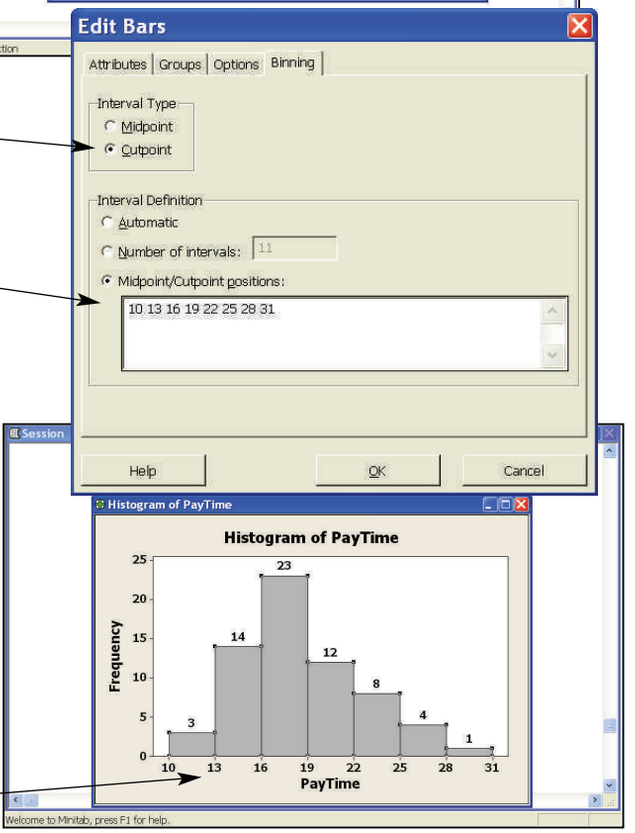
- Right-click inside any of the histogram bars.
- In the pop-up menu, select "Edit bars."



- In the "Edit Bars" dialog box, select the Binning tab.
- To label the *x*-scale by using class boundaries, select the "Interval Type" to be Cutpoint.
- Select the "Interval Definition" to be Midpoint/Cutpoint positions.
- In the Midpoint/Cutpoint positions window, enter the class boundaries (or cutpoints)

  10 13 16 19 22 25 28 31

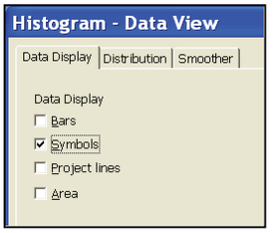
  as given in Table 2.7 or shown in Figure 2.7 (both on page 44).
- If we wished to label the *x*-scale by using class midpoints as in Figure 2.8 on page 45, we would select the "Interval Type" to be Midpoint and we would enter the midpoints of Figure 2.8 (11.5, 14.5, 17.5, and so forth) into the Midpoint/Cutpoint positions window.
- Click OK in the Edit Bars dialog box.



- The histogram in the graphics window will be edited to produce the class boundaries, histogram bars, and *x*-axis labels shown in Figure 2.7.

Frequency Polygons and Ogives: MINITAB does not have automatic procedures for constructing frequency polygons and ogives. However, these graphics can be constructed quite easily by using the MINITAB Graph Annotation Tools. To access these tools and have them placed on the MINITAB toolbar, select

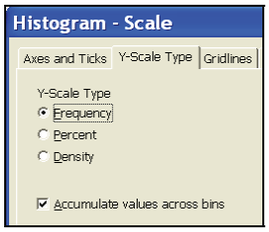
Tools : Toolbars : Graph Annotation Tools



- **To construct a frequency polygon,** follow the preceding instructions for constructing a histogram. In addition, however, click on the Data View button, select the Data Display tab, place a checkmark in the Symbols checkbox (also uncheck the Bars checkbox). This will result in plotted points above the histogram classes—rather than bars. Now select the polygon tool

  from the Graph Annotation Tools toolbar and draw connecting lines to form the polygon. Instructions for using the polygon tool can be found in the MINITAB help resources listed under "To create a polygon."

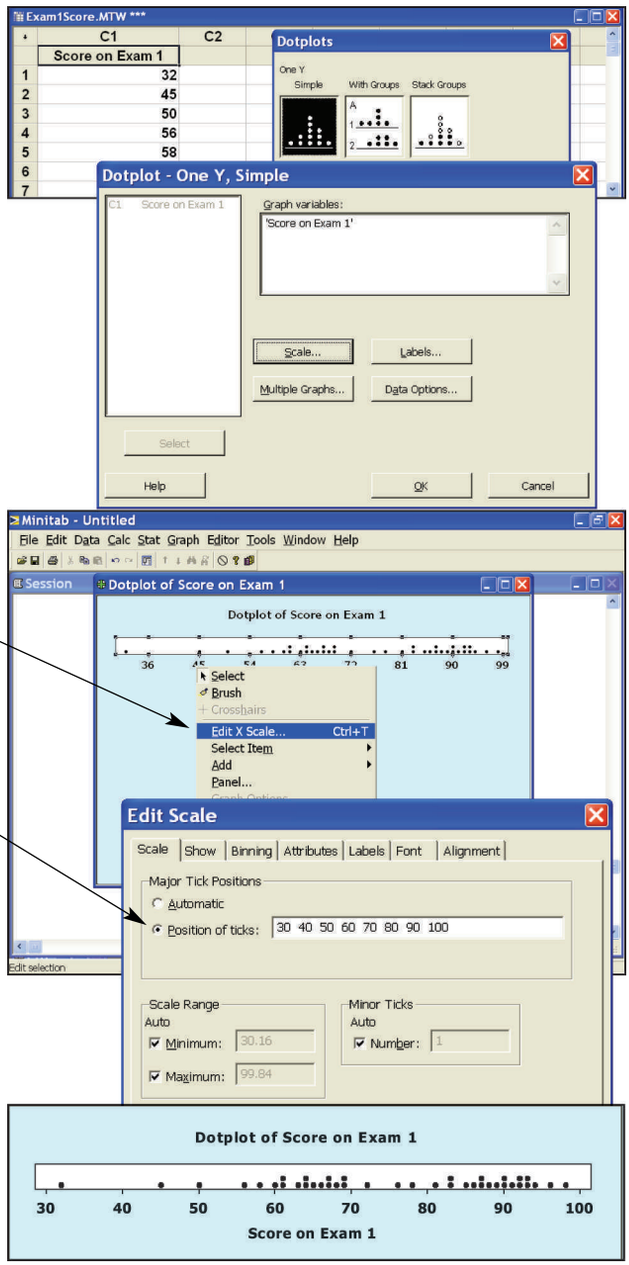
- **To construct an ogive,** follow the above instructions for constructing a frequency polygon. In addition, however, click on the Scale button, select the "Y-Scale Type" tab, and place a checkmark in the "Accumulate values across bins" checkbox. This will result in a plot of cumulative frequencies—rather than histogram bars. Now select the polyline tool



from the Graph Annotation Tools toolbar and draw connecting lines to form the ogive. Instructions for using the polyline tool can be found in the MINITAB help resources listed under "To create a polyline."

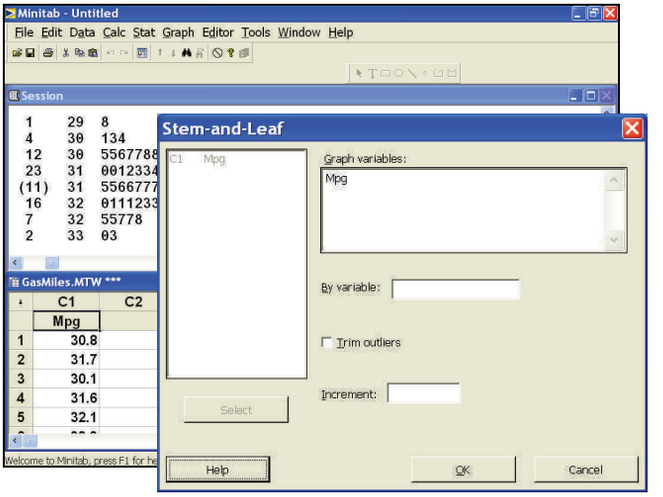
Construct a dot plot of the exam scores as in Figure 2.18(a) on page 55 (data file: FirstExam.MTW):

- Enter the scores for exam 1 in Table 2.8 on page 48 into column C1 with variable name 'Score on Exam 1'.
- Select **Graph : Dot Plot**
- In the Dotplots dialog box, select "One Y Simple" from the gallery of dot plots.
- Click OK in the Dotplots dialog box.
- In the "Dotplot—One Y, Simple" dialog box, enter the variable name 'Score on Exam 1' into the "Graph variables" window. Be sure to include the single quotes.
- Click OK in the "Dotplot—One Y, Simple" dialog box.
- The dotplot will be displayed in a graphics window.
- To change the *x*-axis labels (or, **ticks**), right-click on any one of the existing labels (say, the 45, for instance) and select "Edit X Scale . . ." from the popup menu.
- In the Edit Scale dialog box, select the Scale tab and select "Position of Ticks" as the "Major Tick Positions" setting.
- Enter the desired ticks (30 40 50 60 70 80 90 100) into the "Position of ticks" window and click OK in the Edit Scale dialog box.
- The *x*-axis labels (ticks) will be changed and the new dot plot will be displayed in the graphics window.



Construct a stem-and-leaf display of the gasoline mileages as in Figure 2.19 on page 57 (data file: GasMiles.MTW):

- Enter the mileage data from Table 2.14 on page 56 into column C1 with variable name Mpg.
- Select **Graph : Stem-and-Leaf**
- In the Stem-and-Leaf dialog box, enter the variable name Mpg into the "Graph Variables" window.
- Click OK in the Stem-and-Leaf dialog box.
- The stem-and-leaf display appears in the Session window and can be selected for printing or copied and pasted into a word processing document. (See Appendix 1.3.)



Construct a contingency table of fund type versus level of client satisfaction as in Table 2.17 on page 62 (data file: Invest.MTW):

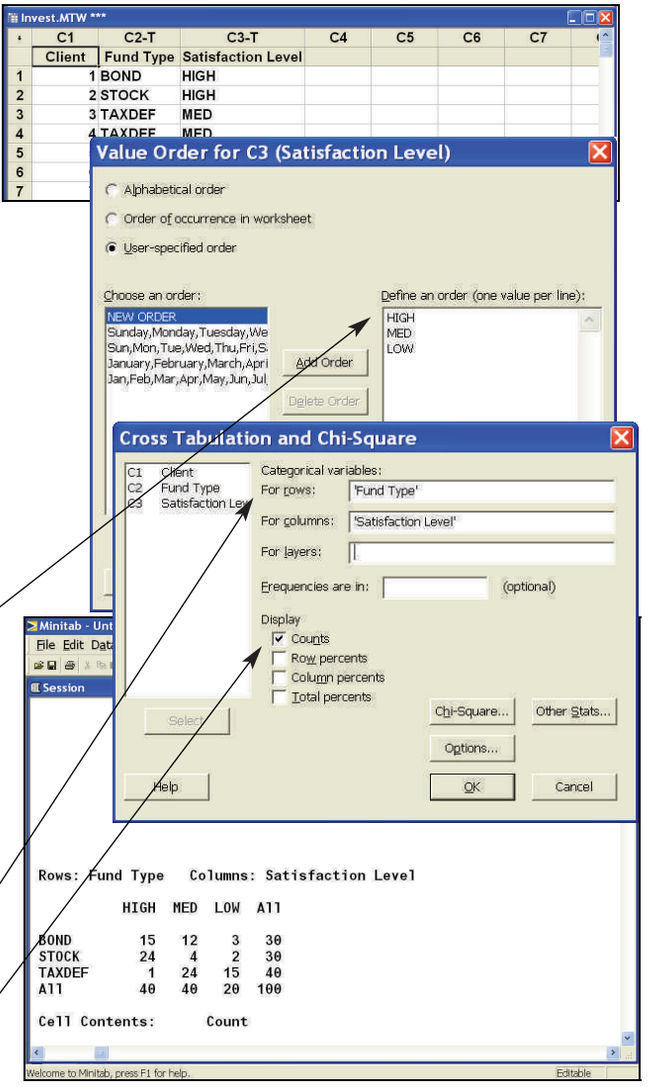
- Enter the client satisfaction data from Table 2.16 on page 62 with client number in column C1 having variable name Client, and with fund type and satisfaction rating in columns C2 and C3, respectively, having variable names 'Fund Type' and 'Satisfaction Level'.

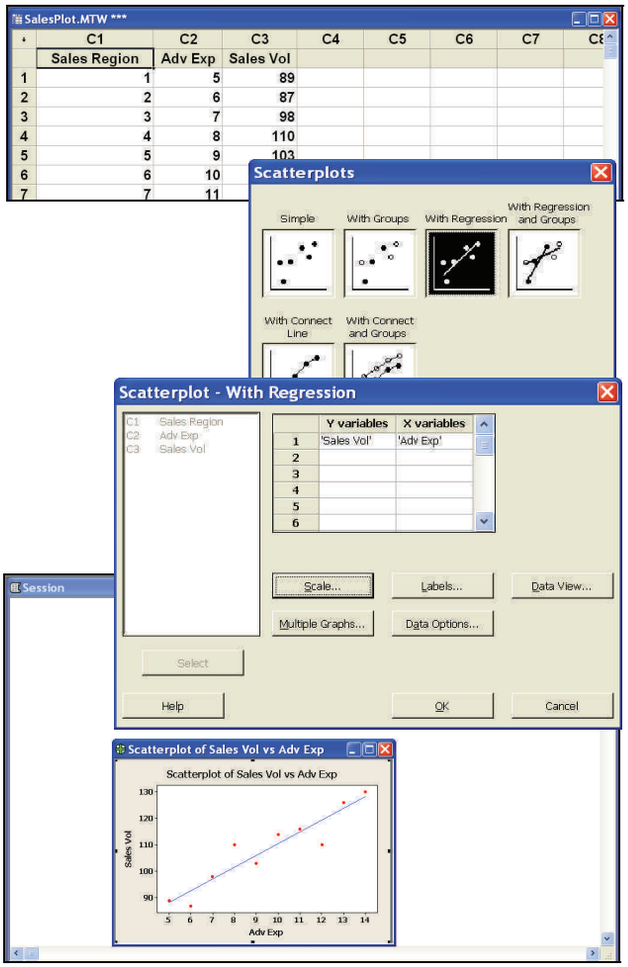
The default ordering for the different levels of each categorical variable in the contingency table will be alphabetical—that is, BOND, STOCK, TAXDEF for 'Fund Type' and HIGH, LOW, MED for 'Satisfaction Level'. **To change the ordering to HIGH, MED, LOW** for 'Satisfaction Level':

- Click on any cell in column C3 (Satisfaction Level).
- Select **Editor : Column : Value order**
- In the "Value Order for C3 (Satisfaction Level)" dialog box, select the "User-specified order" option.
- In the "Define an order (one value per line)" window, specify the order HIGH, MED, LOW.
- Click OK in the "Value Order for C3 (Satisfaction Level)" dialog box.

To construct the contingency table:

- Select **Stat : Tables : Cross Tabulation and Chi-Square**
- In the "Cross Tabulation and Chi-Square" dialog box, enter the variable name 'Fund Type' (including the single quotes) into the "Categorical variables: For rows" window.
- Enter the variable name 'Satisfaction Level' (including the single quotes) into the "Categorical variables: For columns" window.
- Place a checkmark in the "Display Counts" checkbox. We would check "Display Row percents" to produce a table of row percentages and we would check "Display Column percents" to produce a table of column percentages.
- Click OK in the "Cross Tabulation and Chi-Square" dialog box to obtain results in the Session window.



Construct a scatter plot of sales volume versus advertising expenditure as in Figure 2.24 on page 67 (data file: SalesPlot.MWT).

- Enter the sales and advertising data in Table 2.20 (on page 67)—sales region in column C1 (with variable name 'Sales Region'), advertising expenditure in column C2 (with variable name 'Adv Exp'), and sales volume in column C3 (with variable name 'Sales Vol').
- Select **Graph : Scatterplot**
- In the Scatterplots dialog box, select "With Regression" from the gallery of scatterplots in order to produce a scatterplot with a "best line" fitted to the data (see Chapter 14 for discussion of this "best line"). Select "Simple" if a fitted line is not desired.
- Click OK in the Scatterplots dialog box.
- In the "Scatterplot—With Regression" dialog box, enter the variable name 'Sales Vol' (including the single quotes) into row 1 of the "Y variables" window and enter the variable name 'Adv Exp' (including the single quotes) into row 1 of the "X variables" window.
- Click OK in the "Scatterplot—With Regression" dialog box.
- The scatterplot and fitted line will be displayed in a graphics window.
- Additional plots can be obtained by placing appropriate variable names in other rows in the "Y variables" and "X variables" windows.