CHAPTER 2

Descriptive Statistics: Tabular and Graphical Presentations

CONTENTS

STATISTICS *in* PRACTICE

COLGATE-PALMOLIVE COMPANY*

NEW YORK, NEW YORK

The Colgate-Palmolive Company started as a small soap and candle shop in New York City in 1806. Today, Colgate-Palmolive employs more than 40,000 people working in more than 200 countries and territories around the world. Although best known for its brand names of Colgate, Palmolive, Ajax, and Fab, the company also markets Mennen, Hill's Science Diet, and Hill's Prescription Diet products.

The Colgate-Palmolive Company uses statistics in its quality assurance program for home laundry detergent products. One concern is customer satisfaction with the quantity of detergent in a carton. Every carton in each size category is filled with the same amount of detergent by weight, but the volume of detergent is affected by the density of the detergent powder. For instance, if the powder density is on the heavy side, a smaller volume of detergent is needed to reach the carton's specified weight. As a result, the carton may appear to be underfilled when opened by the consumer.

To control the problem of heavy detergent powder, limits are placed on the acceptable range of powder density. Statistical samples are taken periodically, and the density of each powder sample is measured. Data summaries are then provided for operating personnel so that corrective action can be taken if necessary to keep the density within the desired quality specifications.

A frequency distribution for the densities of 150 samples taken over a one-week period and a histogram are shown in the accompanying table and figure. Density levels above .40 are unacceptably high. The frequency distribution and histogram show that the operation is meeting its quality guidelines with all of the densities less than or equal to .40. Managers viewing these statistical summaries would be pleased with the quality of the detergent production process.

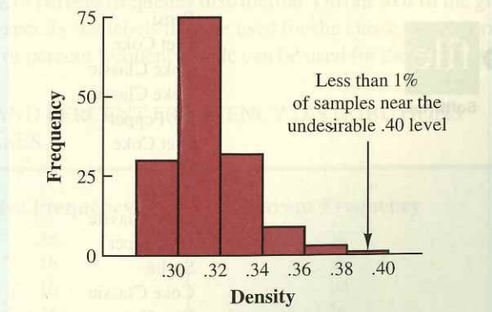
In this chapter, you will learn about tabular and graphical methods of descriptive statistics such as frequency distributions, bar graphs, histograms, stem-and-leaf displays, crosstabulations, and others. The goal of these methods is to summarize data so that the data can be easily understood and interpreted.

Statistical summaries help maintain the quality of these Colgate-Palmolive products. © Joe Higgins/ South-Western.

Frequency Distribution of Density Data

Density	Frequency
.29–.30	30
.31–.32	75
.33–.34	32
.35–.36	9
.37–.38	3
.39–.40	1
Total	150

Histogram of Density Data



*The authors are indebted to William R. Fowle, Manager of Quality Assurance, Colgate-Palmolive Company, for providing this Statistics in Practice.

As indicated in Chapter 1, data can be classified as either qualitative or quantitative. **Qualitative data** use labels or names to identify categories of like items. **Quantitative data** are numerical values that indicate how much or how many.

This chapter introduces tabular and graphical methods commonly used to summarize both qualitative and quantitative data. Tabular and graphical summaries of data can be found in annual reports, newspaper articles, and research studies. Everyone is exposed to these types of presentations. Hence, it is important to understand how they are prepared and how they should be interpreted. We begin with tabular and graphical methods for summarizing data concerning a single variable. The last section introduces methods for summarizing data when the relationship between two variables is of interest.

Modern statistical software packages provide extensive capabilities for summarizing data and preparing graphical presentations. Minitab and Excel are two packages that are widely available. In the chapter appendixes, we show some of their capabilities.

2.1 Summarizing Qualitative Data

Frequency Distribution

We begin the discussion of how tabular and graphical methods can be used to summarize qualitative data with the definition of a **frequency distribution**.

FREQUENCY DISTRIBUTION

A frequency distribution is a tabular summary of data showing the number (frequency) of items in each of several nonoverlapping classes.

Let us use the following example to demonstrate the construction and interpretation of a frequency distribution for qualitative data. Coke Classic, Diet Coke, Dr. Pepper, Pepsi, and Sprite are five popular soft drinks. Assume that the data in Table 2.1 show the soft drink selected in a sample of 50 soft drink purchases.

TABLE 2.1 DATA FROM A SAMPLE OF 50 SOFT DRINK PURCHASES

CD file SoftDrink

Coke Classic	Sprite	Pepsi
Diet Coke	Coke Classic	Coke Classic
Pepsi	Diet Coke	Coke Classic
Diet Coke	Coke Classic	Coke Classic
Coke Classic	Diet Coke	Pepsi
Coke Classic	Coke Classic	Dr. Pepper
Dr. Pepper	Sprite	Coke Classic
Diet Coke	Pepsi	Diet Coke
Pepsi	Coke Classic	Pepsi
Pepsi	Coke Classic	Pepsi
Coke Classic	Coke Classic	Pepsi
Dr. Pepper	Pepsi	Pepsi
Sprite	Coke Classic	Coke Classic
Coke Classic	Sprite	Dr. Pepper
Diet Coke	Dr. Pepper	Pepsi
Coke Classic	Pepsi	Sprite
Coke Classic	Diet Coke	

TABLE 2.2 FREQUENCY DISTRIBUTION OF SOFT DRINK PURCHASES

Soft Drink	Frequency
Coke Classic	19
Diet Coke	8
Dr. Pepper	5
Pepsi	13
Sprite	5
Total	50

To develop a frequency distribution for these data, we count the number of times each soft drink appears in Table 2.1. Coke Classic appears 19 times, Diet Coke appears 8 times, Dr. Pepper appears 5 times, Pepsi appears 13 times, and Sprite appears 5 times. These counts are summarized in the frequency distribution in Table 2.2.

This frequency distribution provides a summary of how the 50 soft drink purchases are distributed across the five soft drinks. This summary offers more insight than the original data shown in Table 2.1. Viewing the frequency distribution, we see that Coke Classic is the leader, Pepsi is second, Diet Coke is third, and Sprite and Dr. Pepper are tied for fourth. The frequency distribution summarizes information about the popularity of the five soft drinks.

Relative Frequency and Percent Frequency Distributions

A frequency distribution shows the number (frequency) of items in each of several nonoverlapping classes. However, we are often interested in the proportion, or percentage, of items in each class. The *relative frequency* of a class equals the fraction or proportion of items belonging to a class. For a data set with n observations, the relative frequency of each class can be determined as follows:

RELATIVE FREQUENCY

$$\text{Relative frequency of a class} = \frac{\text{Frequency of the class}}{n} \quad \textbf{(2.1)}$$

The *percent frequency* of a class is the relative frequency multiplied by 100.

A **relative frequency distribution** gives a tabular summary of data showing the relative frequency for each class. A **percent frequency distribution** summarizes the percent frequency of the data for each class. Table 2.3 shows a relative frequency distribution and a percent frequency distribution for the soft drink data. In Table 2.3 we see that the relative frequency for Coke Classic is 19/50 = .38, the relative frequency for Diet Coke is 8/50 = .16, and so on. From the percent frequency distribution, we see that 38% of the purchases were Coke Classic, 16% of the purchases were Diet Coke, and so on. We can also note that 38% + 26% + 16% = 80% of the purchases were the top three soft drinks.

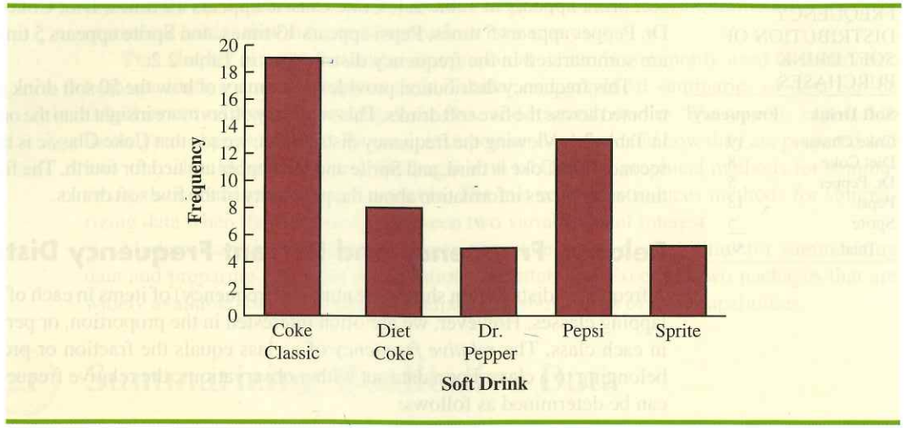
Bar Graphs and Pie Charts

A **bar graph**, or bar chart, is a graphical device for depicting qualitative data summarized in a frequency, relative frequency, or percent frequency distribution. On one axis of the graph (usually the horizontal axis), we specify the labels that are used for the classes (categories). A frequency, relative frequency, or percent frequency scale can be used for the other axis of

TABLE 2.3 RELATIVE FREQUENCY AND PERCENT FREQUENCY DISTRIBUTIONS OF SOFT DRINK PURCHASES

Soft Drink	Relative Frequency	Percent Frequency
Coke Classic	.38	38
Diet Coke	.16	16
Dr. Pepper	.10	10
Pepsi	.26	26
Sprite	.10	10
Total	1.00	100

FIGURE 2.1 BAR GRAPH OF SOFT DRINK PURCHASES

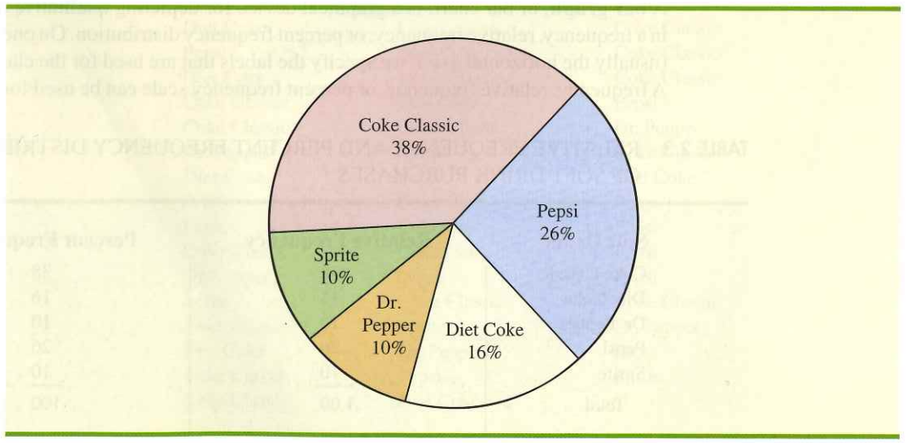


In quality control applications, bar graphs are used to identify the most important causes of problems. When the bars are arranged in descending order of height from left to right with the most frequently occurring cause appearing first, the bar graph is called a pareto diagram. *This diagram is named for its founder, Vilfredo Pareto, an Italian economist.*

the graph (usually the vertical axis). Then, using a bar of fixed width drawn above each class label, we extend the length of the bar until we reach the frequency, relative frequency, or percent frequency of the class. For qualitative data, the bars should be separated to emphasize the fact that each class is separate. Figure 2.1 shows a bar graph of the frequency distribution for the 50 soft drink purchases. Note how the graphical presentation shows Coke Classic, Pepsi, and Diet Coke to be the most preferred brands.

The **pie chart** provides another graphical device for presenting relative frequency and percent frequency distributions for qualitative data. To construct a pie chart, we first draw a circle to represent all of the data. Then we use the relative frequencies to subdivide the circle into sectors, or parts, that correspond to the relative frequency for each class. For example, because a circle contains 360 degrees and Coke Classic shows a relative frequency of .38, the sector of the pie chart labeled Coke Classic consists of .38(360) = 136.8 degrees. The sector of the pie chart labeled Diet Coke consists of .16(360) = 57.6 degrees. Similar calculations for the other classes yield the pie chart in Figure 2.2. The

FIGURE 2.2 PIE CHART OF SOFT DRINK PURCHASES



numerical values shown for each sector can be frequencies, relative frequencies, or percent frequencies.

NOTES AND COMMENTS

1. Often the number of classes in a frequency distribution is the same as the number of categories found in the data, as is the case for the soft drink purchase data in this section. The data involve only five soft drinks, and a separate frequency distribution class was defined for each one. Data that included all soft drinks would require many categories, most of which would have a small number of purchases. Most statisticians recommend that classes with smaller frequencies be grouped into an aggregate class called "other." Classes with frequencies of 5% or less would most often be treated in this fashion.
2. The sum of the frequencies in any frequency distribution always equals the number of observations. The sum of the relative frequencies in any relative frequency distribution always equals 1.00, and the sum of the percentages in a percent frequency distribution always equals 100.

Exercises

Methods

1. The response to a question has three alternatives: A, B, and C. A sample of 120 responses provides 60 A, 24 B, and 36 C. Show the frequency and relative frequency distributions.
2. A partial relative frequency distribution is given.

Class	Relative Frequency
A	.22
B	.18
C	.40
D	

   a. What is the relative frequency of class D?
   b. The total sample size is 200. What is the frequency of class D?
   c. Show the frequency distribution.
   d. Show the percent frequency distribution.

SELF test

3. A questionnaire provides 58 Yes, 42 No, and 20 no-opinion answers.
   a. In the construction of a pie chart, how many degrees would be in the section of the pie showing the Yes answers?
   b. How many degrees would be in the section of the pie showing the No answers?
   c. Construct a pie chart.
   d. Construct a bar graph.

Applications

CD file BestTV

4. The top four primetime television shows were *Law & Order, CSI, Without a Trace,* and *Desperate Housewives* (*Nielsen Media Research,* January 1, 2007). Data indicating the preferred shows for a sample of 50 viewers follow.

DH	*CSI*	*DH*	*CSI*	*L&O*
Trace	*CSI*	*L&O*	*Trace*	*CSI*
CSI	*DH*	*Trace*	*CSI*	*DH*
L&O	*L&O*	*L&O*	*CSI*	*DH*
CSI	*DH*	*DH*	*L&O*	*CSI*
DH	*Trace*	*CSI*	*Trace*	*DH*
DH	*CSI*	*CSI*	*L&O*	*CSI*
L&O	*CSI*	*Trace*	*Trace*	*DH*
L&O	*CSI*	*CSI*	*CSI*	*DH*
CSI	*DH*	*Trace*	*Trace*	*L&O*

a. Are these data qualitative or quantitative?
b. Provide frequency and percent frequency distributions.
c. Construct a bar graph and a pie chart.
d. On the basis of the sample, which television show has the largest viewing audience? Which one is second?

5. In alphabetical order, the six most common last names in the United States are Brown, Davis, Johnson, Jones, Smith, and Williams (*The World Almanac,* 2006). Assume that a sample of 50 individuals with one of these last names provided the following data.

CD file
Names

Brown	Williams	Williams	Williams	Brown
Smith	Jones	Smith	Johnson	Smith
Davis	Smith	Brown	Williams	Johnson
Johnson	Smith	Smith	Johnson	Brown
Williams	Davis	Johnson	Williams	Johnson
Williams	Johnson	Jones	Smith	Brown
Johnson	Smith	Smith	Brown	Jones
Jones	Jones	Smith	Smith	Davis
Davis	Jones	Williams	Davis	Smith
Jones	Johnson	Brown	Johnson	Davis

Summarize the data by constructing the following:
a. Relative and percent frequency distributions
b. A bar graph
c. A pie chart
d. Based on these data, what are the three most common last names?

6. The Nielsen Media Research television rating measures the percentage of television owners who are watching a particular television program. The highest-rated television program in television history was the *M*A*S*H Last Episode Special* shown on February 28, 1983. A 60.2 rating indicated that 60.2% of all television owners were watching this program. Nielsen Media Research provided the list of the 50 top-rated single shows in television history (*The New York Times Almanac,* 2006). The following data show the television network that produced each of these 50 top-rated shows.

CD file
Networks

ABC	ABC	ABC	NBC	CBS
ABC	CBS	ABC	ABC	NBC
NBC	NBC	CBS	ABC	NBC
CBS	ABC	CBS	NBC	ABC
CBS	NBC	NBC	CBS	NBC
CBS	CBS	CBS	NBC	NBC
FOX	CBS	CBS	ABC	NBC
ABC	ABC	CBS	NBC	NBC
NBC	CBS	NBC	CBS	CBS
ABC	CBS	ABC	NBC	ABC

a. Construct a frequency distribution, percent frequency distribution, and bar graph for the data.

b. Which network or networks have done the best in terms of presenting top-rated television shows? Compare the performance of ABC, CBS, and NBC.

SELF test

7. Leverock's Waterfront Steakhouse in Maderia Beach, Florida, uses a questionnaire to ask customers how they rate the server, food quality, cocktails, prices, and atmosphere at the restaurant. Each characteristic is rated on a scale of outstanding (O), very good (V), good (G), average (A), and poor (P). Use descriptive statistics to summarize the following data collected on food quality. What is your feeling about the food quality ratings at the restaurant?

G	O	V	G	A	O	V	O	V	G	O	V	A
V	O	P	V	O	G	A	O	O	O	G	O	V
V	A	G	O	V	P	V	O	O	G	O	O	V
O	G	A	O	V	O	O	G	V	A	G		

8. Data for a sample of 55 members of the Baseball Hall of Fame in Cooperstown, New York, are shown here. Each observation indicates the primary position played by the Hall of Famers: pitcher (P), catcher (H), 1st base (1), 2nd base (2), 3rd base (3), shortstop (S), left field (L), center field (C), and right field (R).

L	P	C	H	2	P	R	1	S	S	1	L	P	R	P
P	P	P	R	C	S	L	R	P	C	C	P	P	R	P
2	3	P	H	L	P	1	C	P	P	P	S	1	L	R
R	1	2	H	S	3	H	2	L	P					

a. Use frequency and relative frequency distributions to summarize the data.
b. What position provides the most Hall of Famers?
c. What position provides the fewest Hall of Famers?
d. What outfield position (L, C, or R) provides the most Hall of Famers?
e. Compare infielders (1, 2, 3, and S) to outfielders (L, C, and R).

9. About 60% of small and medium-sized businesses are family-owned. A TEC International Inc. survey asked the chief executive officers (CEOs) of family-owned businesses how they became the CEO (*The Wall Street Journal,* December 16, 2003). Responses were that the CEO inherited the business, the CEO built the business, or the CEO was hired by the family-owned firm. A sample of 26 CEOs of family-owned businesses provided the following data on how each became the CEO.

CD file
CEOs

Built	Built	Built	Inherited
Inherited	Built	Inherited	Built
Inherited	Built	Built	Built
Built	Hired	Hired	Hired
Inherited	Inherited	Inherited	Built
Built	Built	Built	Hired
Built	Inherited		

a. Provide a frequency distribution.
b. Provide a percent frequency distribution.
c. Construct a bar graph.
d. What percentage of CEOs of family-owned businesses became.the CEO because they inherited the business? What is the primary reason a person becomes the CEO of a family-owned business?

10. Netflix, Inc., of San Jose, California, provides DVD rentals of more than 50,000 titles by mail. Customers go online to create an order list of DVDs they would like to view. Before ordering a particular DVD, the customer may view a description of the DVD and, if desired, a summary of critics' ratings. Netflix uses a five-star rating system with the following descriptions:

1 star	Hated it
2 star	Didn't like it
3 star	Liked it
4 star	Really liked it
5 star	Loved it

Eighteen critics, including Roger Ebert of the *Chicago Sun Times* and Ty Burr of the *Boston Globe,* provided ratings for the movie *Batman Begins* (Netflix.com, March 1, 2006). The ratings for *Batman Begins* were as follows:

4, 2, 5, 2, 4, 3, 3, 4, 4, 3, 4, 4, 4, 2, 4, 4, 5, 4

a. Comment on why these data are qualitative.
b. Provide a frequency distribution and relative frequency distribution for the data.
c. Provide a bar graph.
d. Comment on the critics' evaluation of *Batman Begins.*

2.2 Summarizing Quantitative Data

Frequency Distribution

TABLE 2.4
YEAR-END AUDIT TIMES (IN DAYS)

12	14	19	18
15	15	18	17
20	27	22	23
22	21	33	28
14	18	16	13

As defined in Section 2.1, a frequency distribution is a tabular summary of data showing the number (frequency) of items in each of several nonoverlapping classes. This definition holds for quantitative as well as qualitative data. However, with quantitative data we must be more careful in defining the nonoverlapping classes to be used in the frequency distribution.

For example, consider the quantitative data in Table 2.4. These data show the time in days required to complete year-end audits for a sample of 20 clients of Sanderson and Clifford, a small public accounting firm. The three steps necessary to define the classes for a frequency distribution with quantitative data are:

1. Determine the number of nonoverlapping classes.
2. Determine the width of each class.
3. Determine the class limits.

CD file
Audit

Let us demonstrate these steps by developing a frequency distribution for the audit time data in Table 2.4.

Number of classes Classes are formed by specifying ranges that will be used to group the data. As a general guideline, we recommend using between 5 and 20 classes. For a small number of data items, as few as five or six classes may be used to summarize the data. For a larger number of data items, a larger number of classes is usually required. The goal is to use enough classes to show the variation in the data, but not so many classes that some contain only a few data items. Because the number of data items in Table 2.4 is relatively small ($n = 20$), we chose to develop a frequency distribution with five classes.

Making the classes the same width reduces the chance of inappropriate interpretations by the user.

Width of the classes The second step in constructing a frequency distribution for quantitative data is to choose a width for the classes. As a general guideline, we recommend that the width be the same for each class. Thus the choices of the number of classes and the width of classes are not independent decisions. A larger number of classes means a smaller class width, and vice versa. To determine an approximate class width, we begin by identifying the largest and smallest data values. Then, with the desired number of classes specified, we can use the following expression to determine the approximate class width.

$$\text{Approximate class width} = \frac{\text{Largest data value} - \text{Smallest data value}}{\text{Number of classes}} \qquad (2.2)$$

The approximate class width given by equation (2.2) can be rounded to a more convenient value based on the preference of the person developing the frequency distribution. For example, an approximate class width of 9.28 might be rounded to 10 simply because 10 is a more convenient class width to use in presenting a frequency distribution.

For the data involving the year-end audit times, the largest data value is 33 and the smallest data value is 12. Because we decided to summarize the data with five classes, using

equation (2.2) provides an approximate class width of $(33 - 12)/5 = 4.2$. We therefore decided to round up and use a class width of five days in the frequency distribution.

No single frequency distribution is best for a data set. Different people may construct different, but equally acceptable, frequency distributions. The goal is to reveal the natural grouping and variation in the data.

In practice, the number of classes and the appropriate class width are determined by trial and error. Once a possible number of classes is chosen, equation (2.2) is used to find the approximate class width. The process can be repeated for a different number of classes. Ultimately, the analyst uses judgment to determine the combination of the number of classes and class width that provides the best frequency distribution for summarizing the data.

For the audit time data in Table 2.4, after deciding to use five classes, each with a width of five days, the next task is to specify the class limits for each of the classes.

Class limits Class limits must be chosen so that each data item belongs to one and only one class. The *lower class limit* identifies the smallest possible data value assigned to the class. The *upper class limit* identifies the largest possible data value assigned to the class. In developing frequency distributions for qualitative data, we did not need to specify class limits because each data item naturally fell into a separate class. But with quantitative data, such as the audit times in Table 2.4, class limits are necessary to determine where each data value belongs.

TABLE 2.5
FREQUENCY DISTRIBUTION FOR THE AUDIT TIME DATA

Audit Time (days)	Frequency
10–14	4
15–19	8
20–24	5
25–29	2
30–34	1
Total	20

Using the audit time data in Table 2.4, we selected 10 days as the lower class limit and 14 days as the upper class limit for the first class. This class is denoted 10–14 in Table 2.5. The smallest data value, 12, is included in the 10–14 class. We then selected 15 days as the lower class limit and 19 days as the upper class limit of the next class. We continued defining the lower and upper class limits to obtain a total of five classes: 10–14, 15–19, 20–24, 25–29, and 30–34. The largest data value, 33, is included in the 30–34 class. The difference between the lower class limits of adjacent classes is the class width. Using the first two lower class limits of 10 and 15, we see that the class width is $15 - 10 = 5$.

With the number of classes, class width, and class limits determined, a frequency distribution can be obtained by counting the number of data values belonging to each class. For example, the data in Table 2.4 show that four values—12, 14, 14, and 13—belong to the 10–14 class. Thus, the frequency for the 10–14 class is 4. Continuing this counting process for the 15–19, 20–24, 25–29, and 30–34 classes provides the frequency distribution in Table 2.5. Using this frequency distribution, we can observe the following:

1. The most frequently occurring audit times are in the class of 15–19 days. Eight of the 20 audit times belong to this class.
2. Only one audit required 30 or more days.

Other conclusions are possible, depending on the interests of the person viewing the frequency distribution. The value of a frequency distribution is that it provides insights about the data that are not easily obtained by viewing the data in their original unorganized form.

Class midpoint In some applications, we want to know the midpoints of the classes in a frequency distribution for quantitative data. The **class midpoint** is the value halfway between the lower and upper class limits. For the audit time data, the five class midpoints are 12, 17, 22, 27, and 32.

Relative Frequency and Percent Frequency Distributions

We define the relative frequency and percent frequency distributions for quantitative data in the same manner as for qualitative data. First, recall that the relative frequency is the proportion of the observations belonging to a class. With n observations,

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

The percent frequency of a class is the relative frequency multiplied by 100.

Based on the class frequencies in Table 2.5 and with $n = 20$, Table 2.6 shows the relative frequency distribution and percent frequency distribution for the audit time data. Note that .40

TABLE 2.6 RELATIVE FREQUENCY AND PERCENT FREQUENCY DISTRIBUTIONS FOR THE AUDIT TIME DATA

Audit Time (days)	Relative Frequency	Percent Frequency
10–14	.20	20
15–19	.40	40
20–24	.25	25
25–29	.10	10
30–34	.05	5
Total	1.00	100

of the audits, or 40%, required from 15 to 19 days. Only .05 of the audits, or 5%, required 30 or more days. Again, additional interpretations and insights can be obtained by using Table 2.6.

Dot Plot

One of the simplest graphical summaries of data is a **dot plot**. A horizontal axis shows the range for the data. Each data value is represented by a dot placed above the axis. Figure 2.3 is the dot plot for the audit time data in Table 2.4. The three dots located above 18 on the horizontal axis indicate that an audit time of 18 days occurred three times. Dot plots show the details of the data and are useful for comparing the distribution of the data for two or more variables.

Histogram

A common graphical presentation of quantitative data is a **histogram**. This graphical summary can be prepared for data previously summarized in either a frequency, relative frequency, or percent frequency distribution. A histogram is constructed by placing the variable of interest on the horizontal axis and the frequency, relative frequency, or percent frequency on the vertical axis. The frequency, relative frequency, or percent frequency of each class is shown by drawing a rectangle whose base is determined by the class limits on the horizontal axis and whose height is the corresponding frequency, relative frequency, or percent frequency.

Figure 2.4 is a histogram for the audit time data. Note that the class with the greatest frequency is shown by the rectangle appearing above the class of 15–19 days. The height of the rectangle shows that the frequency of this class is 8. A histogram for the relative or percent frequency distribution of these data would look the same as the histogram in Figure 2.4 with the exception that the vertical axis would be labeled with relative or percent frequency values.

As Figure 2.4 shows, the adjacent rectangles of a histogram touch one another. Unlike a bar graph, a histogram contains no natural separation between the rectangles of adjacent classes. This format is the usual convention for histograms. Because the classes for the audit

FIGURE 2.3 DOT PLOT FOR THE AUDIT TIME DATA

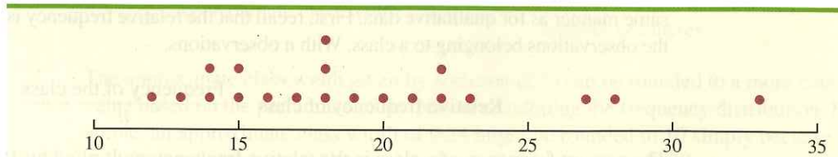
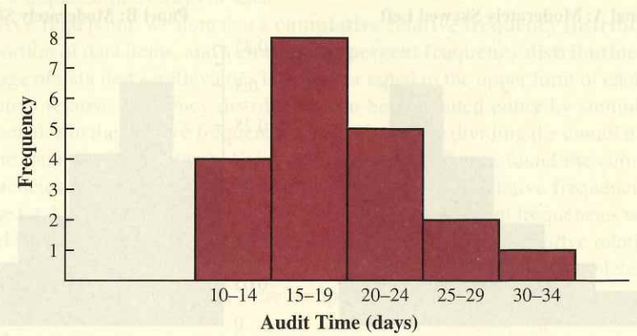


FIGURE 2.4 HISTOGRAM FOR THE AUDIT TIME DATA



time data are stated as 10–14, 15–19, 20–24, 25–29, and 30–34, one-unit spaces of 14 to 15, 19 to 20, 24 to 25, and 29 to 30 would seem to be needed between the classes. These spaces are eliminated when constructing a histogram. Eliminating the spaces between classes in a histogram for the audit time data helps show that all values between the lower limit of the first class and the upper limit of the last class are possible.

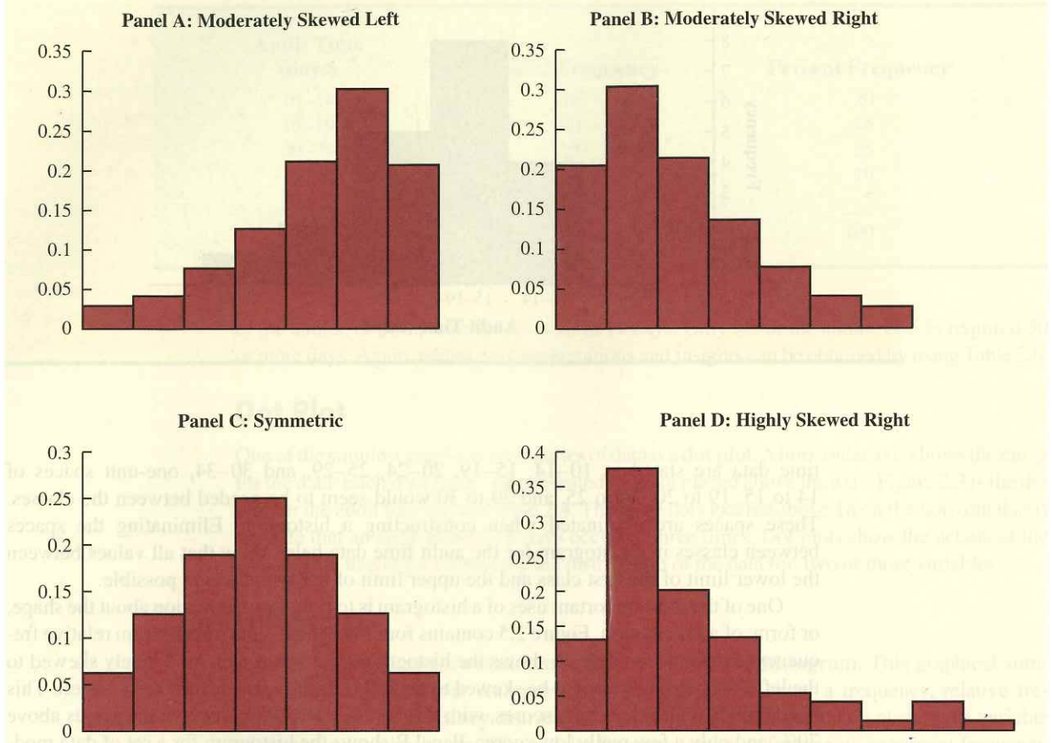
One of the most important uses of a histogram is to provide information about the shape, or form, of a distribution. Figure 2.5 contains four histograms constructed from relative frequency distributions. Panel A shows the histogram for a set of data moderately skewed to the left. A histogram is said to be skewed to the left if its tail extends farther to the left. This histogram is typical for exam scores, with no scores above 100%, most of the scores above 70%, and only a few really low scores. Panel B shows the histogram for a set of data moderately skewed to the right. A histogram is said to be skewed to the right if its tail extends farther to the right. An example of this type of histogram would be for data such as housing prices; a few expensive houses create the skewness in the right tail.

Panel C shows a symmetric histogram. In a symmetric histogram, the left tail mirrors the shape of the right tail. Histograms for data found in applications are never perfectly symmetric, but the histogram for many applications may be roughly symmetric. Data for SAT scores, heights and weights of people, and so on lead to histograms that are roughly symmetric. Panel D shows a histogram highly skewed to the right. This histogram was constructed from data on the amount of customer purchases over one day at a women's apparel store. Data from applications in business and economics often lead to histograms that are skewed to the right. For instance, data on housing prices, salaries, purchase amounts, and so on often result in histograms skewed to the right.

Cumulative Distributions

A variation of the frequency distribution that provides another tabular summary of quantitative data is the **cumulative frequency distribution**. The cumulative frequency distribution uses the number of classes, class widths, and class limits developed for the frequency distribution. However, rather than showing the frequency of each class, the cumulative frequency distribution shows the number of data items with values *less than or equal to the upper class limit* of each class. The first two columns of Table 2.7 provide the cumulative frequency distribution for the audit time data.

FIGURE 2.5 HISTOGRAMS SHOWING DIFFERING LEVELS OF SKEWNESS



To understand how the cumulative frequencies are determined, consider the class with the description "less than or equal to 24." The cumulative frequency for this class is simply the sum of the frequencies for all classes with data values less than or equal to 24. For the frequency distribution in Table 2.5, the sum of the frequencies for classes 10–14, 15–19, and 20–24 indicates that 4 + 8 + 5 = 17 data values are less than or equal to 24. Hence,

TABLE 2.7 CUMULATIVE FREQUENCY, CUMULATIVE RELATIVE FREQUENCY, AND CUMULATIVE PERCENT FREQUENCY DISTRIBUTIONS FOR THE AUDIT TIME DATA

Audit Time (days)	Cumulative Frequency	Cumulative Relative Frequency	Cumulative Percent Frequency
Less than or equal to 14	4	.20	20
Less than or equal to 19	12	.60	60
Less than or equal to 24	17	.85	85
Less than or equal to 29	19	.95	95
Less than or equal to 34	20	1.00	100

the cumulative frequency for this class is 17. In addition, the cumulative frequency distribution in Table 2.7 shows that four audits were completed in 14 days or less and 19 audits were completed in 29 days or less.

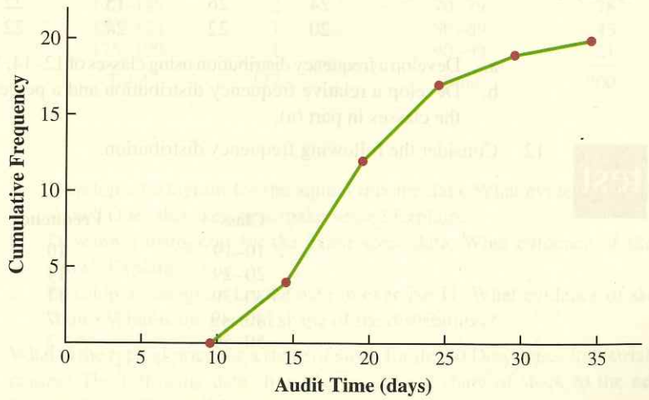
As a final point, we note that a **cumulative relative frequency distribution** shows the proportion of data items, and a **cumulative percent frequency distribution** shows the percentage of data items with values less than or equal to the upper limit of each class. The cumulative relative frequency distribution can be computed either by summing the relative frequencies in the relative frequency distribution or by dividing the cumulative frequencies by the total number of items. Using the latter approach, we found the cumulative relative frequencies in column 3 of Table 2.7 by dividing the cumulative frequencies in column 2 by the total number of items ($n = 20$). The cumulative percent frequencies were again computed by multiplying the relative frequencies by 100. The cumulative relative and percent frequency distributions show that .85 of the audits, or 85%, were completed in 24 days or less, .95 of the audits, or 95%, were completed in 29 days or less, and so on.

Ogive

A graph of a cumulative distribution, called an **ogive**, shows data values on the horizontal axis and either the cumulative frequencies, the cumulative relative frequencies, or the cumulative percent frequencies on the vertical axis. Figure 2.6 illustrates an ogive for the cumulative frequencies of the audit time data in Table 2.7.

The ogive is constructed by plotting a point corresponding to the cumulative frequency of each class. Because the classes for the audit time data are 10–14, 15–19, 20–24, and so on, one-unit gaps appear from 14 to 15, 19 to 20, and so on. These gaps are eliminated by plotting points halfway between the class limits. Thus, 14.5 is used for the 10–14 class, 19.5 is used for the 15–19 class, and so on. The "less than or equal to 14" class with a cumulative frequency of 4 is shown on the ogive in Figure 2.6 by the point located at 14.5 on the horizontal axis and 4 on the vertical axis. The "less than or equal to 19" class with a cumulative frequency of 12 is shown by the point located at 19.5 on the horizontal axis and 12 on the vertical axis. Note that one additional point is plotted at the left end of the ogive. This point starts the ogive by showing that no data values fall below the 10–14 class. It is plotted at 9.5 on the horizontal axis and 0 on the vertical axis. The plotted points are connected by straight lines to complete the ogive.

FIGURE 2.6 OGIVE FOR THE AUDIT TIME DATA



NOTES AND COMMENTS

1. A bar graph and a histogram are essentially the same thing; both are graphical presentations of the data in a frequency distribution. A histogram is just a bar graph with no separation between bars. For some discrete quantitative data, a separation between bars is also appropriate. Consider, for example, the number of classes in which a college student is enrolled. The data may only assume integer values. Intermediate values such as 1.5, 2.73, and so on are not possible. With continuous quantitative data, however, such as the audit times in Table 2.4, a separation between bars is not appropriate.
2. The appropriate values for the class limits with quantitative data depend on the level of accuracy of the data. For instance, with the audit time data of Table 2.4 the limits used were integer values. If the data were rounded to the nearest tenth of a day (e.g., 12.3, 14.4, and so on), then the limits would be stated in tenths of days. For instance, the first class would be 10.0–14.9. If the data were recorded to the nearest hundredth of a day (e.g., 12.34, 14.45, and so on), the limits would be stated in hundredths of days. For instance, the first class would be 10.00–14.99.
3. An *open-end* class requires only a lower class limit or an upper class limit. For example, in the audit time data of Table 2.4, suppose two of the audits had taken 58 and 65 days. Rather than continue with the classes of width 5 with classes 35–39, 40–44, 45–49, and so on, we could simplify the frequency distribution to show an open-end class of "35 or more." This class would have a frequency of 2. Most often the open-end class appears at the upper end of the distribution. Sometimes an open-end class appears at the lower end of the distribution, and occasionally such classes appear at both ends.
4. The last entry in a cumulative frequency distribution always equals the total number of observations. The last entry in a cumulative relative frequency distribution always equals 1.00 and the last entry in a cumulative percent frequency distribution always equals 100.

Exercises

Methods

11. Consider the following data.

CD file Frequency

14	21	23	21	16
19	22	25	16	16
24	24	25	19	16
19	18	19	21	12
16	17	18	23	25
20	23	16	20	19
24	26	15	22	24
20	22	24	22	20

a. Develop a frequency distribution using classes of 12–14, 15–17, 18–20, 21–23, and 24–26.
b. Develop a relative frequency distribution and a percent frequency distribution using the classes in part (a).

SELF test

12. Consider the following frequency distribution.

Class	Frequency
10–19	10
20–29	14
30–39	17
40–49	7
50–59	2

Construct a cumulative frequency distribution and a cumulative relative frequency distribution.

13. Construct a histogram and an ogive for the data in exercise 12.
14. Consider the following data.

8.9	10.2	11.5	7.8	10.0	12.2	13.5	14.1	10.0	12.2
6.8	9.5	11.5	11.2	14.9	7.5	10.0	6.0	15.8	11.5

a. Construct a dot plot.
b. Construct a frequency distribution.
c. Construct a percent frequency distribution.

Applications

SELF test

15. A doctor's office staff studied the waiting times for patients who arrive at the office with a request for emergency service. The following data with waiting times in minutes were collected over a one-month period.

2 5 10 12 4 4 5 17 11 8 9 8 12 21 6 8 7 13 18 3

Use classes of 0–4, 5–9, and so on in the following:
a. Show the frequency distribution.
b. Show the relative frequency distribution.
c. Show the cumulative frequency distribution.
d. Show the cumulative relative frequency distribution.
e. What proportion of patients needing emergency service wait 9 minutes or less?

16. Consider the following two frequency distributions. The first frequency distribution provides an approximation of the annual adjusted gross income in the United States (Internal Revenue Service, March 2003). The second frequency distribution shows exam scores for students in a college statistics course.

Income ($1000s)	Frequency (millions)
0–24	60
25–49	33
50–74	20
75–99	6
100–124	4
125–149	2
150–174	1
175–199	1
Total	127

Exam Score	Frequency
20–29	2
30–39	5
40–49	6
50–59	13
60–69	32
70–79	78
80–89	43
90–99	21
Total	200

a. Develop a histogram for the annual income data. What evidence of skewness does it show? Does this skewness make sense? Explain.
b. Develop a histogram for the exam score data. What evidence of skewness does it show? Explain.
c. Develop a histogram for the data in exercise 11. What evidence of skewness does it show? What is the general shape of the distribution?

17. What is the typical price for a share of stock for the 30 Dow Jones Industrial Average companies? The following data show the price for a share of stock to the nearest dollar in January 2006 (*The Wall Street Journal,* January 16, 2006).

CD file PriceShare

Company	$/Share	Company	$/Share
AIG	70	Home Depot	42
Alcoa	29	Honeywell	37
Altria Group	76	IBM	83
American Express	53	Intel	26
AT&T	25	Johnson & Johnson	62
Boeing	69	JPMorgan Chase	40
Caterpillar	62	McDonald's	35
Citigroup	49	Merck	33
Coca-Cola	41	Microsoft	27
Disney	26	3M	78
DuPont	40	Pfizer	25
ExxonMobil	61	Procter & Gamble	59
General Electric	35	United Technologies	56
General Motors	20	Verizon	32
Hewlett-Packard	32	Wal-Mart	45

a. Prepare a frequency distribution of the data.
b. Prepare a histogram of the data. Interpret the histogram, including a discussion of the general shape of the histogram, the mid-price per share range, the most frequent price per share range, and the high and low extreme prices per share.
c. What are the highest-priced and the lowest-priced stocks?
d. Use *The Wall Street Journal* to find the current price per share for these companies. Prepare a histogram of the data and discuss any changes since January 2006.

18. NRF/BIG research provided results of a consumer holiday spending survey (*USA Today,* December 20, 2005). The following data provide the dollar amount of holiday spending for a sample of 25 consumers.

CD file Holiday

1200	850	740	590	340
450	890	260	610	350
1780	180	850	2050	770
800	1090	510	520	220
1450	280	1120	200	350

a. What is the lowest holiday spending? The highest?
b. Use a class width of $250 to prepare a frequency distribution and a percent frequency distribution for the data.
c. Prepare a histogram and comment on the shape of the distribution.
d. What observations can you make about holiday spending?

19. Sorting through unsolicited e-mail and spam affects the productivity of office workers. An InsightExpress survey monitored office workers to determine the unproductive time per day devoted to unsolicited e-mail and spam (*USA Today,* November 13, 2003). The following data show a sample of time in minutes devoted to this task.

2	4	8	4
8	1	2	32
12	1	5	7
5	5	3	4
24	19	4	14

Summarize the data by constructing the following:
a. A frequency distribution (classes 1–5, 6–10, 11–15, 16–20, and so on)
b. A relative frequency distribution
c. A cumulative frequency distribution

d. A cumulative relative frequency distribution
e. An ogive
f. What percentage of office workers spend 5 minutes or less on unsolicited e-mail and spam? What percentage of office workers spend more than 10 minutes a day on this task?

20. The top 20 concert tours and their average ticket price for shows in North America are shown here. The list is based on data provided to the trade publication *Pollstar* by concert promoters and venue managers (*Associated Press,* November 21, 2003).

CD file Concerts

Concert Tour	Ticket Price	Concert Tour	Ticket Price
Bruce Springsteen	$72.40	Toby Keith	$37.76
Dave Matthews Band	44.11	James Taylor	44.93
Aerosmith/KISS	69.52	Alabama	40.83
Shania Twain	61.80	Harper/Johnson	33.70
Fleetwood Mac	78.34	50 Cent	38.89
Radiohead	39.50	Steely Dan	36.38
Cher	64.47	Red Hot Chili Peppers	56.82
Counting Crows	36.48	R.E.M.	46.16
Timberlake/Aguilera	74.43	American Idols Live	39.11
Mana	46.48	Mariah Carey	56.08

Summarize the data by constructing the following:
a. A frequency distribution and a percent frequency distribution
b. A histogram
c. What concert had the most expensive average ticket price? What concert had the least expensive average ticket price?
d. Comment on what the data indicate about the average ticket prices of the top concert tours.

21. The *Nielsen Home Technology Report* provided information about home technology and its usage. The following data are the hours of personal computer usage during one week for a sample of 50 persons.

CD file Computer

4.1	1.5	10.4	5.9	3.4	5.7	1.6	6.1	3.0	3.7
3.1	4.8	2.0	14.8	5.4	4.2	3.9	4.1	11.1	3.5
4.1	4.1	8.8	5.6	4.3	3.3	7.1	10.3	6.2	7.6
10.8	2.8	9.5	12.9	12.1	0.7	4.0	9.2	4.4	5.7
7.2	6.1	5.7	5.9	4.7	3.9	3.7	3.1	6.1	3.1

Summarize the data by constructing the following:
a. A frequency distribution (use a class width of three hours)
b. A relative frequency distribution
c. A histogram
d. An ogive
e. Comment on what the data indicate about personal computer usage at home.

2.3 Exploratory Data Analysis: The Stem-and-Leaf Display

The techniques of **exploratory data analysis** consist of simple arithmetic and easy-to-draw graphs that can be used to summarize data quickly. One technique—referred to as a **stem-and-leaf display**—can be used to show both the rank order and shape of a data set simultaneously.

TABLE 2.8 NUMBER OF QUESTIONS ANSWERED CORRECTLY ON AN APTITUDE TEST

112	72	69	97	107
73	92	76	86	73
126	128	118	127	124
82	104	132	134	83
92	108	96	100	92
115	76	91	102	81
95	141	81	80	106
84	119	113	98	75
68	98	115	106	95
100	85	94	106	119

CD file ApTest

To illustrate the use of a stem-and-leaf display, consider the data in Table 2.8. These data result from a 150-question aptitude test given to 50 individuals recently interviewed for a position at Haskens Manufacturing. The data indicate the number of questions answered correctly.

To develop a stem-and-leaf display, we first arrange the leading digits of each data value to the left of a vertical line. To the right of the vertical line, we record the last digit for each data value. Based on the top row of data in Table 2.8 (112, 72, 69, 97, and 107), the first five entries in constructing a stem-and-leaf display would be as follows:

6	9
7	2
8	
9	7
10	7
11	2
12	
13	
14	

For example, the data value 112 shows the leading digits 11 to the left of the line and the last digit 2 to the right of the line. Similarly, the data value 72 shows the leading digit 7 to the left of the line and last digit 2 to the right of the line. Continuing to place the last digit of each data value on the line corresponding to its leading digit(s) provides the following:

6	9 8
7	2 3 6 3 6 5
8	6 2 3 1 1 0 4 5
9	7 2 2 6 2 1 5 8 8 5 4
10	7 4 8 0 2 6 6 0 6
11	2 8 5 9 3 5 9
12	6 8 7 4
13	2 4
14	1

With this organization of the data, sorting the digits on each line into rank order is simple. Doing so provides the stem-and-leaf display shown here.

6	8 9
7	2 3 3 5 6 6
8	0 1 1 2 3 4 5 6
9	1 2 2 2 4 5 5 6 7 8 8
10	0 0 2 4 6 6 6 7 8
11	2 3 5 5 8 9 9
12	4 6 7 8
13	2 4
14	1

The numbers to the left of the vertical line (6, 7, 8, 9, 10, 11, 12, 13, and 14) form the *stem,* and each digit to the right of the vertical line is a *leaf.* For example, consider the first row with a stem value of 6 and leaves of 8 and 9.

6 | 8 9

This row indicates that two data values have a first digit of six. The leaves show that the data values are 68 and 69. Similarly, the second row

7 | 2 3 3 5 6 6

indicates that six data values have a first digit of seven. The leaves show that the data values are 72, 73, 73, 75, 76, and 76.

To focus on the shape indicated by the stem-and-leaf display, let us use a rectangle to contain the leaves of each stem. Doing so, we obtain the following.

6	8 9
7	2 3 3 5 6 6
8	0 1 1 2 3 4 5 6
9	1 2 2 2 4 5 5 6 7 8 8
10	0 0 2 4 6 6 6 7 8
11	2 3 5 5 8 9 9
12	4 6 7 8
13	2 4
14	1

Rotating this page counterclockwise onto its side provides a picture of the data that is similar to a histogram with classes of 60–69, 70–79, 80–89, and so on.

Although the stem-and-leaf display may appear to offer the same information as a histogram, it has two primary advantages.

1. The stem-and-leaf display is easier to construct by hand.
2. Within a class interval, the stem-and-leaf display provides more information than the histogram because the stem-and-leaf shows the actual data.

Just as a frequency distribution or histogram has no absolute number of classes, neither does a stem-and-leaf display have an absolute number of rows or stems. If we believe that our original stem-and-leaf display condensed the data too much, we can easily stretch the display by using two or more stems for each leading digit. For example, to use two stems for each leading digit,

In a stretched stem-and-leaf display, whenever a stem value is stated twice, the first value corresponds to leaf values of 0–4, and the second value corresponds to leaf values of 5–9.

we would place all data values ending in 0, 1, 2, 3, and 4 in one row and all values ending in 5, 6, 7, 8, and 9 in a second row. The following stretched stem-and-leaf display illustrates this approach.

```
 6 | 8 9
 7 | 2 3 3
 7 | 5 6 6
 8 | 0 1 1 2 3 4
 8 | 5 6
 9 | 1 2 2 2 4
 9 | 5 5 6 7 8 8
10 | 0 0 2 4
10 | 6 6 6 7 8
11 | 2 3
11 | 5 5 8 9 9
12 | 4
12 | 6 7 8
13 | 2 4
13 |
14 | 1
```

Note that values 72, 73, and 73 have leaves in the 0–4 range and are shown with the first stem value of 7. The values 75, 76, and 76 have leaves in the 5–9 range and are shown with the second stem value of 7. This stretched stem-and-leaf display is similar to a frequency distribution with intervals of 65–69, 70–74, 75–79, and so on.

The preceding example showed a stem-and-leaf display for data with as many as three digits. Stem-and-leaf displays for data with more than three digits are possible. For example, consider the following data on the number of hamburgers sold by a fast-food restaurant for each of 15 weeks.

1565	1852	1644	1766	1888	1912	2044	1812
1790	1679	2008	1852	1967	1954	1733	

A stem-and-leaf display of these data follows.

```
Leaf unit = 10

15 | 6
16 | 4 7
17 | 3 6 9
18 | 1 5 5 8
19 | 1 5 6
20 | 0 4
```

A single digit is used to define each leaf in a stem-and-leaf display. The leaf unit indicates how to multiply the stem-and-leaf numbers in order to approximate the original data. Leaf units may be 100, 10, 1, 0.1, and so on.

Note that a single digit is used to define each leaf and that only the first three digits of each data value have been used to construct the display. At the top of the display we have specified Leaf unit = 10. To illustrate how to interpret the values in the display, consider the first stem, 15, and its associated leaf, 6. Combining these numbers, we obtain 156. To reconstruct an approximation of the original data value, we must multiply this number by 10, the value of the *leaf unit*. Thus, $156 \times 10 = 1560$ is an approximation of the original data value used to construct the stem-and-leaf display. Although it is not possible to reconstruct the exact data value from this stem-and-leaf display, the convention of using a single digit for each leaf enables stem-and-leaf displays to be constructed for data having a large number of digits. For stem-and-leaf displays where the leaf unit is not shown, the leaf unit is assumed to equal 1.

Exercises

Methods

22. Construct a stem-and-leaf display for the following data.

70	72	75	64	58	83	80	82
76	75	68	65	57	78	85	72

SELF test

23. Construct a stem-and-leaf display for the following data.

11.3	9.6	10.4	7.5	8.3	10.5	10.0
9.3	8.1	7.7	7.5	8.4	6.3	8.8

24. Construct a stem-and-leaf display for the following data. Use a leaf unit of 10.

1161	1206	1478	1300	1604	1725	1361	1422
1221	1378	1623	1426	1557	1730	1706	1689

Applications

SELF test

25. A psychologist developed a new test of adult intelligence. The test was administered to 20 individuals, and the following data were obtained.

114	99	131	124	117	102	106	127	119	115
98	104	144	151	132	106	125	122	118	118

Construct a stem-and-leaf display for the data.

26. The American Association of Individual Investors conducts an annual survey of discount brokers. The following prices charged are from a sample of 24 discount brokers (*AAII Journal,* January 2003). The two types of trades are a broker-assisted trade of 100 shares at $50 per share and an online trade of 500 shares at $50 per share.

CD file Broker

Broker	Broker-Assisted 100 Shares at $50/Share	Online 500 Shares at $50/Share	Broker	Broker-Assisted 100 Shares at $50/Share	Online 500 Shares at $50/Share
Accutrade	30.00	29.95	Merrill Lynch Direct	50.00	29.95
Ameritrade	24.99	10.99	Muriel Siebert	45.00	14.95
Banc of America	54.00	24.95	NetVest	24.00	14.00
Brown & Co.	17.00	5.00	Recom Securities	35.00	12.95
Charles Schwab	55.00	29.95	Scottrade	17.00	7.00
CyberTrader	12.95	9.95	Sloan Securities	39.95	19.95
E*TRADE Securities	49.95	14.95	Strong Investments	55.00	24.95
First Discount	35.00	19.75	TD Waterhouse	45.00	17.95
Freedom Investments	25.00	15.00	T. Rowe Price	50.00	19.95
Harrisdirect	40.00	20.00	Vanguard	48.00	20.00
Investors National	39.00	62.50	Wall Street Discount	29.95	19.95
MB Trading	9.95	10.55	York Securities	40.00	36.00

a. Round the trading prices to the nearest dollar and develop a stem-and-leaf display for 100 shares at $50 per share. Comment on what you learned about broker-assisted trading prices.

b. Round the trading prices to the nearest dollar and develop a stretched stem-and-leaf display for 500 shares online at $50 per share. Comment on what you learned about online trading prices.

27. Most major ski resorts offer family programs that provide ski and snowboarding instruction for children. The typical classes provide four to six hours on the snow with a certified instructor. The daily rate for a group lesson at 15 ski resorts follows (*The Wall Street Journal,* January 20, 2006).

Resort	Location	Daily Rate	Resort	Location	Daily Rate
Beaver Creek	Colorado	$137	Okemo	Vermont	$ 86
Deer Valley	Utah	115	Park City	Utah	145
Diamond Peak	California	95	Butternut	Massachusetts	75
Heavenly	California	145	Steamboat	Colorado	98
Hunter	New York	79	Stowe	Vermont	104
Mammoth	California	111	Sugar Bowl	California	100
Mount Sunapee	New Hampshire	96	Whistler-Blackcomb	British Columbia	104
Mount Bachelor	Oregon	83			

a. Develop a stem-and-leaf display for the data.
b. Interpret the stem-and-leaf display in terms of what it tells you about the daily rate for these ski and snowboarding instruction programs.

28. The 2004 Naples, Florida, mini marathon (13.1 miles) had 1228 registrants (*Naples Daily News,* January 17, 2004). Competition was held in six age groups. The following data show the ages for a sample of 40 individuals who participated in the marathon.

CD file
Marathon

49	33	40	37	56
44	46	57	55	32
50	52	43	64	40
46	24	30	37	43
31	43	50	36	61
27	44	35	31	43
52	43	66	31	50
72	26	59	21	47

a. Show a stretched stem-and-leaf display.
b. What age group had the largest number of runners?
c. What age occurred most frequently?
d. A *Naples Daily News* feature article emphasized the number of runners who were "20-something." What percentage of the runners were in the 20-something age group? What do you suppose was the focus of the article?

2.4 Crosstabulations and Scatter Diagrams

Crosstabulations and scatter diagrams are used to summarize data in a way that reveals the relationship between two variables.

Thus far in this chapter, we have focused on tabular and graphical methods used to summarize the data for *one variable at a time.* Often a manager or decision maker requires tabular and graphical methods that will assist in the understanding of the *relationship between two variables.* Crosstabulation and scatter diagrams are two such methods.

Crosstabulation

A **crosstabulation** is a tabular summary of data for two variables. Let us illustrate the use of a crosstabulation by considering the following application based on data from Zagat's Restaurant Review. The quality rating and the meal price data were collected for a sample of 300 restaurants located in the Los Angeles area. Table 2.9 shows the data for the first 10 restaurants. Data on a restaurant's quality rating and typical meal price are reported. Quality rating is a qualitative variable with rating categories of good, very good, and excellent. Meal price is a quantitative variable that ranges from $10 to $49.

A crosstabulation of the data for this application is shown in Table 2.10. The left and top margin labels define the classes for the two variables. In the left margin, the row labels (good, very good, and excellent) correspond to the three classes of the quality rating variable. In the top margin, the column labels ($10–19, $20–29, $30–39, and $40–49) correspond to

TABLE 2.9 QUALITY RATING AND MEAL PRICE FOR 300 LOS ANGELES RESTAURANTS

CD file
Restaurant

Restaurant	Quality Rating	Meal Price ($)
1	Good	18
2	Very Good	22
3	Good	28
4	Excellent	38
5	Very Good	33
6	Good	28
7	Very Good	19
8	Very Good	11
9	Very Good	23
10	Good	13
.	.	.
.	.	.
.	.	.

the four classes of the meal price variable. Each restaurant in the sample provides a quality rating and a meal price. Thus, each restaurant in the sample is associated with a cell appearing in one of the rows and one of the columns of the crosstabulation. For example, restaurant 5 is identified as having a very good quality rating and a meal price of $33. This restaurant belongs to the cell in row 2 and column 3 of Table 2.10. In constructing a crosstabulation, we simply count the number of restaurants that belong to each of the cells in the crosstabulation table.

In reviewing Table 2.10, we see that the greatest number of restaurants in the sample (64) have a very good rating and a meal price in the $20–29 range. Only two restaurants have an excellent rating and a meal price in the $10–19 range. Similar interpretations of the other frequencies can be made. In addition, note that the right and bottom margins of the crosstabulation provide the frequency distributions for quality rating and meal price separately. From the frequency distribution in the right margin, we see that data on quality ratings show 84 good restaurants, 150 very good restaurants, and 66 excellent restaurants. Similarly, the bottom margin shows the frequency distribution for the meal price variable.

Dividing the totals in the right margin of the crosstabulation by the total for that column provides a relative and percent frequency distribution for the quality rating variable.

Quality Rating	Relative Frequency	Percent Frequency
Good	.28	28
Very Good	.50	50
Excellent	.22	22
Total	1.00	100

TABLE 2.10 CROSSTABULATION OF QUALITY RATING AND MEAL PRICE FOR 300 LOS ANGELES RESTAURANTS

	Meal Price				
Quality Rating	$10–19	$20–29	$30–39	$40–49	Total
Good	42	40	2	0	84
Very Good	34	64	46	6	150
Excellent	2	14	28	22	66
Total	78	118	76	28	300

From the percent frequency distribution we see that 28% of the restaurants were rated good, 50% were rated very good, and 22% were rated excellent.

Dividing the totals in the bottom row of the crosstabulation by the total for that row provides a relative and percent frequency distribution for the meal price variable.

Meal Price	Relative Frequency	Percent Frequency
$10–19	.26	26
$20–29	.39	39
$30–39	.25	25
$40–49	.09	9
Total	1.00	100

Note that the sum of the values in each column does not add exactly to the column total, because the values being summed are rounded. From the percent frequency distribution we see that 26% of the meal prices are in the lowest price class ($10–19), 39% are in the next higher class, and so on.

The frequency and relative frequency distributions constructed from the margins of a crosstabulation provide information about each of the variables individually, but they do not shed any light on the relationship between the variables. The primary value of a crosstabulation lies in the insight it offers about the relationship between the variables. A review of the crosstabulation in Table 2.10 reveals that higher meal prices are associated with the higher quality restaurants, and the lower meal prices are associated with the lower quality restaurants.

Converting the entries in a crosstabulation into row percentages or column percentages can provide more insight into the relationship between the two variables. For row percentages, the results of dividing each frequency in Table 2.10 by its corresponding row total are shown in Table 2.11. Each row of Table 2.11 is a percent frequency distribution of meal price for one of the quality rating categories. Of the restaurants with the lowest quality rating (good), we see that the greatest percentages are for the less expensive restaurants (50% have $10–19 meal prices and 47.6% have $20–29 meal prices). Of the restaurants with the highest quality rating (excellent), we see that the greatest percentages are for the more expensive restaurants (42.4% have $30–39 meal prices and 33.4% have $40–49 meal prices). Thus, we continue to see that the more expensive meals are associated with the higher quality restaurants.

Crosstabulation is widely used for examining the relationship between two variables. In practice, the final reports for many statistical studies include a large number of crosstabulation tables. In the Los Angeles restaurant survey, the crosstabulation is based on one qualitative variable (quality rating) and one quantitative variable (meal price). Crosstabulations can also be developed when both variables are qualitative and when both variables are quantitative. When quantitative variables are used, however, we must first create classes for the values of the variable. For instance, in the restaurant example we grouped the meal prices into four classes ($10–19, $20–29, $30–39, and $40–49).

TABLE 2.11 ROW PERCENTAGES FOR EACH QUALITY RATING CATEGORY

	Meal Price				
Quality Rating	$10–19	$20–29	$30–39	$40–49	Total
Good	50.0	47.6	2.4	0.0	100
Very Good	22.7	42.7	30.6	4.0	100
Excellent	3.0	21.2	42.4	33.4	100

Simpson's Paradox

The data in two or more crosstabulations are often combined or aggregated to produce a summary crosstabulation showing how two variables are related. In such cases, we must be careful in drawing conclusions about the relationship between the two variables in the aggregated crosstabulation. In some cases the conclusions based upon the aggregated crosstabulation can be completely reversed if we look at the unaggregated data, an occurrence known as **Simpson's paradox**. To provide an illustration of Simpson's paradox we consider an example involving the analysis of verdicts for two judges in two types of courts.

Judges Ron Luckett and Dennis Kendall presided over cases in Common Pleas Court and Municipal Court during the past three years. Some of the verdicts they rendered were appealed. In most of these cases the appeals court upheld the original verdicts, but in some cases those verdicts were reversed. For each judge a crosstabulation was developed based upon two variables: Verdict (upheld or reversed) and Type of Court (Common Pleas and Municipal). Suppose that the two crosstabulations were then combined by aggregating the type of court data. The resulting aggregated crosstabulation contains two variables: Verdict (upheld or reversed) and Judge (Luckett or Kendall). This crosstabulation shows the number of appeals in which the verdict was upheld and the number in which the verdict was reversed for both judges. The following crosstabulation shows these results along with the column percentages in parentheses next to each value.

	Judge		
Verdict	Luckett	Kendall	Total
Upheld	129 (86%)	110 (88%)	239
Reversed	21 (14%)	15 (12%)	36
Total (%)	150 (100%)	125 (100%)	275

A review of the column percentages shows that 14% of the verdicts were reversed for Judge Luckett, but only 12% of the verdicts were reversed for Judge Kendall. Thus, we might conclude that Judge Kendall is doing a better job because a higher percentage of his verdicts are being upheld. A problem arises with this conclusion, however.

The following crosstabulations show the cases tried by Luckett and Kendall in the two courts; column percentages are also shown in parentheses next to each value.

	Judge Luckett		
Verdict	Common Pleas	Municipal Court	Total
Upheld	29 (91%)	100 (85%)	129
Reversed	3 (9%)	18 (15%)	21
Total (%)	32 (100%)	118 (100%)	150

	Judge Kendall		
Verdict	Common Pleas	Municipal Court	Total
Upheld	90 (90%)	20 (80%)	110
Reversed	10 (10%)	5 (20%)	15
Total (%)	100 (100%)	25 (100%)	125

From the crosstabulation and column percentages for Luckett, we see that his verdicts were upheld in 91% of the Common Pleas Court cases and in 85% of the Municipal Court cases. From the crosstabulation and column percentages for Kendall, we see that his verdicts were upheld in 90% of the Common Pleas Court cases and in 80% of the Municipal Court cases. Comparing the column percentages for the two judges, we see that Judge Luckett demonstrates a better record than Judge Kendall in both courts. This result contradicts the conclusion we reached when we aggregated the data across both courts for the original crosstabulation. It appeared then that Judge Kendall had the better record. This example illustrates Simpson's paradox.

The original crosstabulation was obtained by aggregating the data in the separate crosstabulations for the two courts. Note that for both judges the percentage of appeals that resulted in reversals was much higher in Municipal Court than in Common Pleas Court. Because Judge Luckett tried a much higher percentage of his cases in Municipal Court, the aggregated data favored Judge Kendall. When we look at the crosstabulations for the two courts separately, however, Judge Luckett clearly shows the better record. Thus, for the original crosstabulation, we see that the *type of court* is a hidden variable that cannot be ignored when evaluating the records of the two judges.

Because of Simpson's paradox, we need to be especially careful when drawing conclusions using aggregated data. Before drawing any conclusions about the relationship between two variables shown for a crosstabulation involving aggregated data, you should investigate whether any hidden variables could affect the results.

Scatter Diagram and Trendline

A **scatter diagram** is a graphical presentation of the relationship between two quantitative variables, and a **trendline** is a line that provides an approximation of the relationship. As an illustration, consider the advertising/sales relationship for a stereo and sound equipment store in San Francisco. On 10 occasions during the past three months, the store used weekend television commercials to promote sales at its stores. The managers want to investigate whether a relationship exists between the number of commercials shown and sales at the store during the following week. Sample data for the 10 weeks with sales in hundreds of dollars are shown in Table 2.12.

Figure 2.7 shows the scatter diagram and the trendline* for the data in Table 2.12. The number of commercials (x) is shown on the horizontal axis and the sales (y) are shown on the vertical axis. For week 1, $x = 2$ and $y = 50$. A point with those coordinates is plotted on the scatter diagram. Similar points are plotted for the other nine weeks. Note that during two of the weeks one commercial was shown, during two of the weeks two commercials were shown, and so on.

The completed scatter diagram in Figure 2.7 indicates a positive relationship between the number of commercials and sales. Higher sales are associated with a higher number of commercials. The relationship is not perfect in that all points are not on a straight line. However, the general pattern of the points and the trendline suggest that the overall relationship is positive.

Some general scatter diagram patterns and the types of relationships they suggest are shown in Figure 2.8. The top left panel depicts a positive relationship similar to the one for

TABLE 2.12 SAMPLE DATA FOR THE STEREO AND SOUND EQUIPMENT STORE

CD file Stereo

Week	Number of Commercials x	Sales ($100s) y
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

*The equation of the trendline is $y = 36.15 + 4.95x$. The slope of the trendline is 4.95 and the y-intercept (the point where the line intersects the y axis) is 36.15. We will discuss in detail the interpretation of the slope and y-intercept for a linear trendline in Chapter 12 when we study simple linear regression.

FIGURE 2.7 SCATTER DIAGRAM AND TRENDLINE FOR THE STEREO AND SOUND EQUIPMENT STORE

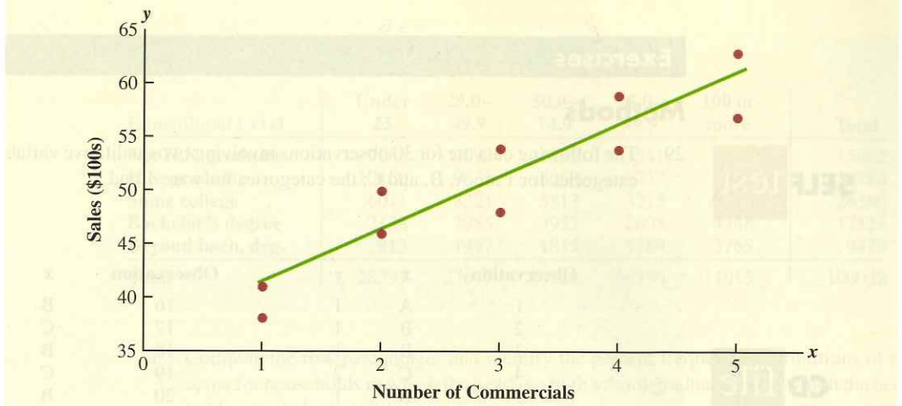
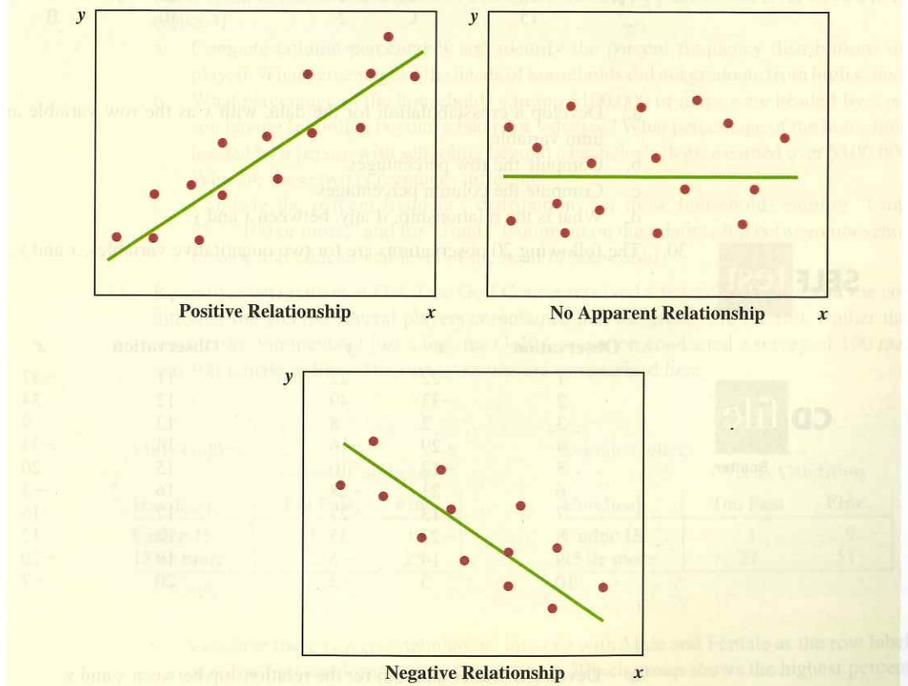


FIGURE 2.8 TYPES OF RELATIONSHIPS DEPICTED BY SCATTER DIAGRAMS



the number of commercials and sales example. In the top right panel, the scatter diagram shows no apparent relationship between the variables. The bottom panel depicts a negative relationship where y tends to decrease as x increases.

Exercises

Methods

SELF test

29. The following data are for 30 observations involving two qualitative variables, x and y. The categories for x are A, B, and C; the categories for y are 1 and 2.

CD file
Crosstab

Observation	x	y	Observation	x	y
1	A	1	16	B	2
2	B	1	17	C	1
3	B	1	18	B	1
4	C	2	19	C	1
5	B	1	20	B	1
6	C	2	21	C	2
7	B	1	22	B	1
8	C	2	23	C	2
9	A	1	24	A	1
10	B	1	25	B	1
11	A	1	26	C	2
12	B	1	27	C	2
13	C	2	28	A	1
14	C	2	29	B	1
15	C	2	30	B	2

a. Develop a crosstabulation for the data, with x as the row variable and y as the column variable.
b. Compute the row percentages.
c. Compute the column percentages.
d. What is the relationship, if any, between x and y?

SELF test

30. The following 20 observations are for two quantitative variables, x and y.

CD file
Scatter

Observation	x	y	Observation	x	y
1	−22	22	11	−37	48
2	−33	49	12	34	−29
3	2	8	13	9	−18
4	29	−16	14	−33	31
5	−13	10	15	20	−16
6	21	−28	16	−3	14
7	−13	27	17	−15	18
8	−23	35	18	12	17
9	14	−5	19	−20	−11
10	3	−3	20	−7	−22

a. Develop a scatter diagram for the relationship between x and y.
b. What is the relationship, if any, between x and y?

Applications

31. The following crosstabulation shows household income by educational level of the head of household (*Statistical Abstract of the United States: 2002*).

	Household Income ($1000s)					
Educational Level	Under 25	25.0–49.9	50.0–74.9	75.0–99.9	100 or more	Total
Not H.S. graduate	9285	4093	1589	541	354	15862
H.S. graduate	10150	9821	6050	2737	2028	30786
Some college	6011	8221	5813	3215	3120	26380
Bachelor's degree	2138	3985	3952	2698	4748	17521
Beyond bach. deg.	813	1497	1815	1589	3765	9479
Total	28397	27617	19219	10780	14015	100028

a. Compute the row percentages and identify the percent frequency distributions of income for households in which the head is a high school graduate and in which the head holds a bachelor's degree.
b. What percentage of households headed by high school graduates earn $75,000 or more? What percentage of households headed by bachelor's degree recipients earn $75,000 or more?
c. Construct percent frequency histograms of income for households headed by persons with a high school degree and for those headed by persons with a bachelor's degree. Is any relationship evident between household income and educational level?

32. Refer again to the crosstabulation of household income by educational level shown in exercise 31.
a. Compute column percentages and identify the percent frequency distributions displayed. What percentage of the heads of households did not graduate from high school?
b. What percentage of the households earning $100,000 or more were headed by a person having schooling beyond a bachelor's degree? What percentage of the households headed by a person with schooling beyond a bachelor's degree earned over $100,000? Why are these two percentages different?
c. Compare the percent frequency distributions for those households earning "Under 25," "100 or more," and for "Total." Comment on the relationship between household income and educational level of the head of household.

33. Recently, management at Oak Tree Golf Course received a few complaints about the condition of the greens. Several players complained that the greens are too fast. Rather than react to the comments of just a few, the Golf Association conducted a survey of 100 male and 100 female golfers. The survey results are summarized here.

Male Golfers

	Greens Condition	
Handicap	Too Fast	Fine
Under 15	10	40
15 or more	25	25

Female Golfers

	Greens Condition	
Handicap	Too Fast	Fine
Under 15	1	9
15 or more	39	51

a. Combine these two crosstabulations into one with Male and Female as the row labels and Too Fast and Fine as the column labels. Which group shows the highest percentage saying that the greens are too fast?

b. Refer to the initial crosstabulations. For those players with low handicaps (better players), which group (male or female) shows the highest percentage saying the greens are too fast?
c. Refer to the initial crosstabulations. For those players with higher handicaps, which group (male or female) shows the highest percentage saying the greens are too fast?
d. What conclusions can you draw about the preferences of men and women concerning the speed of the greens? Are the conclusions you draw from part (a) as compared with parts (b) and (c) consistent? Explain any apparent inconsistencies.

34. Table 2.13 provides financial data for a sample of 36 companies whose stocks trade on the New York Stock Exchange (*Investor's Business Daily,* April 7, 2000). The data on Sales/Margins/ROE are a composite rating based on a company's sales growth rate, its profit margins, and its return on equity (ROE). EPS Rating is a measure of growth in earnings per share for the company.

TABLE 2.13 FINANCIAL DATA FOR A SAMPLE OF 36 COMPANIES

CD file IBD

Company	EPS Rating	Relative Price Strength	Industry Group Relative Strength	Sales/Margins/ ROE
Advo	81	74	B	A
Alaska Air Group	58	17	C	B
Alliant Tech	84	22	B	B
Atmos Energy	21	9	C	E
Bank of Am.	87	38	C	A
Bowater PLC	14	46	C	D
Callaway Golf	46	62	B	E
Central Parking	76	18	B	C
Dean Foods	84	7	B	C
Dole Food	70	54	E	C
Elec. Data Sys.	72	69	A	B
Fed. Dept. Store	79	21	D	B
Gateway	82	68	A	A
Goodyear	21	9	E	D
Hanson PLC	57	32	B	B
ICN Pharm.	76	56	A	D
Jefferson Plt.	80	38	D	C
Kroger	84	24	D	A
Mattel	18	20	E	D
McDermott	6	6	A	C
Monaco	97	21	D	A
Murphy Oil	80	62	B	B
Nordstrom	58	57	B	C
NYMAGIC	17	45	D	D
Office Depot	58	40	B	B
Payless Shoes	76	59	B	B
Praxair	62	32	C	B
Reebok	31	72	C	E
Safeway	91	61	D	A
Teco Energy	49	48	D	B
Texaco	80	31	D	C
US West	60	65	B	A
United Rental	98	12	C	A
Wachovia	69	36	E	B
Winnebago	83	49	D	A
York International	28	14	D	B

Source: Investor's Business Daily, April 7, 2000.

a. Prepare a crosstabulation of the data on Sales/Margins/ROE (rows) and EPS Rating (columns). Use classes of 0–19, 20–39, 40–59, 60–79, and 80–99 for EPS Rating.
b. Compute row percentages and comment on any relationship between the variables.

35. Refer to the data in Table 2.13.
a. Prepare a crosstabulation of the data on Sales/Margins/ROE and Industry Group Relative Strength.
b. Prepare a frequency distribution for the data on Sales/Margins/ROE.
c. Prepare a frequency distribution for the data on Industry Group Relative Strength.
d. How has the crosstabulation helped in preparing the frequency distributions in parts (b) and (c)?

36. Refer to the data in Table 2.13.
a. Prepare a scatter diagram of the data on EPS Rating and Relative Price Strength.
b. Comment on the relationship, if any, between the variables. (The meaning of the EPS Rating is described in exercise 34. Relative Price Strength is a measure of the change in the stock's price over the past 12 months. Higher values indicate greater strength.)

37. The National Football League rates prospects position by position on a scale that ranges from 5 to 9. The ratings are interpreted as follows: 8–9 should start the first year; 7.0–7.9 should start; 6.0–6.9 will make the team as a backup; and 5.0–5.9 can make the club and contribute. Table 2.14 shows the position, weight, time (seconds to run 40 yards), and rating for 40 NFL prospects (*USA Today,* April 14, 2000).
a. Prepare a crosstabulation of the data on Position (rows) and Time (columns). Use classes of 4.00–4.49, 4.50–4.99, 5.00–5.49, and 5.50–5.99 for Time.
b. Comment on the relationship between Position and Time based upon the crosstabulation developed in part (a).
c. Develop a scatter diagram of the data on Time and Rating. Use the vertical axis for Rating.
d. Comment on the relationship, if any, between Time and Rating.

Summary

A set of data, even if modest in size, is often difficult to interpret directly in the form in which it is gathered. Tabular and graphical methods provide procedures for organizing and summarizing data so that patterns are revealed and the data are more easily interpreted. Frequency distributions, relative frequency distributions, percent frequency distributions, bar graphs, and pie charts were presented as tabular and graphical procedures for summarizing qualitative data. Frequency distributions, relative frequency distributions, percent frequency distributions, histograms, cumulative frequency distributions, cumulative relative frequency distributions, cumulative percent frequency distributions, and ogives were presented as ways of summarizing quantitative data. A stem-and-leaf display provides an exploratory data analysis technique that can be used to summarize quantitative data. Crosstabulation was presented as a tabular method for summarizing data for two variables. The scatter diagram was introduced as a graphical method for showing the relationship between two quantitative variables. Figure 2.9 shows the tabular and graphical methods presented in this chapter.

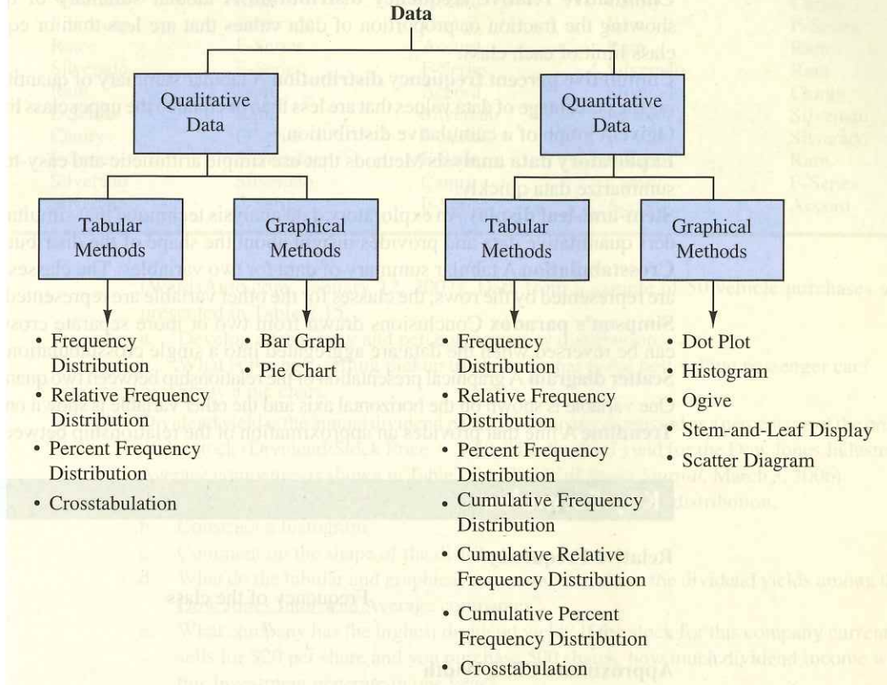
With large data sets, computer software packages are essential in constructing tabular and graphical summaries of data. In the two chapter appendixes, we show how Minitab and Excel can be used for this purpose.

TABLE 2.14 NATIONAL FOOTBALL LEAGUE RATINGS FOR 40 DRAFT PROSPECTS

CD file NFL

Observation	Name	Position	Weight	Time	Rating
1	Peter Warrick	Wide receiver	194	4.53	9
2	Plaxico Burress	Wide receiver	231	4.52	8.8
3	Sylvester Morris	Wide receiver	216	4.59	8.3
4	Travis Taylor	Wide receiver	199	4.36	8.1
5	Laveranues Coles	Wide receiver	192	4.29	8
6	Dez White	Wide receiver	218	4.49	7.9
7	Jerry Porter	Wide receiver	221	4.55	7.4
8	Ron Dugans	Wide receiver	206	4.47	7.1
9	Todd Pinkston	Wide receiver	169	4.37	7
10	Dennis Northcutt	Wide receiver	175	4.43	7
11	Anthony Lucas	Wide receiver	194	4.51	6.9
12	Darrell Jackson	Wide receiver	197	4.56	6.6
13	Danny Farmer	Wide receiver	217	4.6	6.5
14	Sherrod Gideon	Wide receiver	173	4.57	6.4
15	Trevor Gaylor	Wide receiver	199	4.57	6.2
16	Cosey Coleman	Guard	322	5.38	7.4
17	Travis Claridge	Guard	303	5.18	7
18	Kaulana Noa	Guard	317	5.34	6.8
19	Leander Jordan	Guard	330	5.46	6.7
20	Chad Clifton	Guard	334	5.18	6.3
21	Manula Savea	Guard	308	5.32	6.1
22	Ryan Johanningmeir	Guard	310	5.28	6
23	Mark Tauscher	Guard	318	5.37	6
24	Blaine Saipaia	Guard	321	5.25	6
25	Richard Mercier	Guard	295	5.34	5.8
26	Damion McIntosh	Guard	328	5.31	5.3
27	Jeno James	Guard	320	5.64	5
28	Al Jackson	Guard	304	5.2	5
29	Chris Samuels	Offensive tackle	325	4.95	8.5
30	Stockar McDougle	Offensive tackle	361	5.5	8
31	Chris McIngosh	Offensive tackle	315	5.39	7.8
32	Adrian Klemm	Offensive tackle	307	4.98	7.6
33	Todd Wade	Offensive tackle	326	5.2	7.3
34	Marvel Smith	Offensive tackle	320	5.36	7.1
35	Michael Thompson	Offensive tackle	287	5.05	6.8
36	Bobby Williams	Offensive tackle	332	5.26	6.8
37	Darnell Alford	Offensive tackle	334	5.55	6.4
38	Terrance Beadles	Offensive tackle	312	5.15	6.3
39	Tutan Reyes	Offensive tackle	299	5.35	6.1
40	Greg Robinson-Ran	Offensive tackle	333	5.59	6

FIGURE 2.9 TABULAR AND GRAPHICAL METHODS FOR SUMMARIZING DATA



Glossary

Qualitative data Labels or names used to identify categories of like items.
Quantitative data Numerical values that indicate how much or how many.
Frequency distribution A tabular summary of data showing the number (frequency) of data values in each of several nonoverlapping classes.
Relative frequency distribution A tabular summary of data showing the fraction or proportion of data values in each of several nonoverlapping classes.
Percent frequency distribution A tabular summary of data showing the percentage of data values in each of several nonoverlapping classes.
Bar graph A graphical device for depicting qualitative data that have been summarized in a frequency, relative frequency, or percent frequency distribution.
Pie chart A graphical device for presenting data summaries based on subdivision of a circle into sectors that correspond to the relative frequency for each class.
Class midpoint The value halfway between the lower and upper class limits.
Dot plot A graphical device that summarizes data by the number of dots above each data value on the horizontal axis.
Histogram A graphical presentation of a frequency distribution, relative frequency distribution, or percent frequency distribution of quantitative data constructed by placing the class intervals on the horizontal axis and the frequencies, relative frequencies, or percent frequencies on the vertical axis.

Cumulative frequency distribution A tabular summary of quantitative data showing the number of data values that are less than or equal to the upper class limit of each class.
Cumulative relative frequency distribution A tabular summary of quantitative data showing the fraction or proportion of data values that are less than or equal to the upper class limit of each class.
Cumulative percent frequency distribution A tabular summary of quantitative data showing the percentage of data values that are less than or equal to the upper class limit of each class.
Ogive A graph of a cumulative distribution.
Exploratory data analysis Methods that use simple arithmetic and easy-to-draw graphs to summarize data quickly.
Stem-and-leaf display An exploratory data analysis technique that simultaneously rank orders quantitative data and provides insight about the shape of the distribution.
Crosstabulation A tabular summary of data for two variables. The classes for one variable are represented by the rows; the classes for the other variable are represented by the columns.
Simpson's paradox Conclusions drawn from two or more separate crosstabulations that can be reversed when the data are aggregated into a single crosstabulation.
Scatter diagram A graphical presentation of the relationship between two quantitative variables. One variable is shown on the horizontal axis and the other variable is shown on the vertical axis.
Trendline A line that provides an approximation of the relationship between two variables.

Key Formulas

Relative Frequency

$$\frac{\text{Frequency of the class}}{n} \tag{2.1}$$

Approximate Class Width

$$\frac{\text{Largest data value} - \text{Smallest data value}}{\text{Number of classes}} \tag{2.2}$$

Supplementary Exercises

38. The Higher Education Research Institute at UCLA provides statistics on the most popular majors among incoming college freshmen. The five most popular majors are Arts and Humanities (A), Business Administration (B), Engineering (E), Professional (P), and Social Science (S) (*The New York Times Almanac,* 2006). A broad range of other (O) majors, including biological science, physical science, computer science, and education, are grouped together. The majors selected for a sample of 64 college freshmen follow.

CD file Major

S	P	P	O	B	E	O	E	P	O	O	B	O	O	O	A
O	E	E	B	S	O	B	O	A	O	E	O	E	O	B	P
B	A	S	O	E	A	B	O	S	S	O	O	E	B	O	B
A	E	B	E	A	A	P	O	O	E	O	B	B	O	P	B

a. Show a frequency distribution and percent frequency distribution.
b. Show a bar graph.
c. What percentage of freshmen selects one of the five most popular majors?
d. What is the most popular major for incoming freshmen? What percentage of freshmen select this major?

39. Five of the top-selling vehicles during 2006 were the Chevrolet Silverado/C/K pickup, Dodge Ram pickup, Ford F-Series pickup, Honda Accord, and Toyota Camry

TABLE 2.15 DATA FOR 50 VEHICLE PURCHASES

CD file AutoData

Silverado	Ram	Accord	Camry	Camry
Silverado	Silverado	Camry	Ram	F-Series
Ram	F-Series	Accord	Ram	Ram
Silverado	F-Series	F-Series	Silverado	Ram
Ram	Ram	Accord	Silverado	Camry
F-Series	Ram	Silverado	Accord	Silverado
Camry	F-Series	F-Series	F-Series	Silverado
F-Series	Silverado	F-Series	F-Series	Ram
Silverado	Silverado	Camry	Camry	F-Series
Silverado	F-Series	F-Series	Accord	Accord

(WardsAuto.com, January 12, 2007). Data from a sample of 50 vehicle purchases are presented in Table 2.15.
a. Develop a frequency and percent frequency distribution.
b. What is the best-selling pickup truck, and what is the best-selling passenger car?
c. Show a pie chart.

40. Dividend yield is the annual dividend paid by a company expressed as a percentage of the price of the stock (Dividend/Stock Price × 100). The dividend yield for the Dow Jones Industrial Average companies is shown in Table 2.16 (*The Wall Street Journal,* March 3, 2006).
a. Construct a frequency distribution and percent frequency distribution.
b. Construct a histogram.
c. Comment on the shape of the distribution.
d. What do the tabular and graphical summaries tell about the dividend yields among the Dow Jones Industrial Average companies?
e. What company has the highest dividend yield? If the stock for this company currently sells for $20 per share and you purchase 500 shares, how much dividend income will this investment generate in one year?

41. *Golf Magazine*'s Top 100 Teachers were asked the question, "What is the most critical area that prevents golfers from reaching their potential?" The possible responses were lack of accuracy, poor approach shots, poor mental approach, lack of power, limited practice, poor

TABLE 2.16 DIVIDEND YIELD FOR DOW JONES INDUSTRIAL AVERAGE COMPANIES

CD file DivYield

Company	Dividend Yield %	Company	Dividend Yield %
AIG	0.9	Home Depot	1.4
Alcoa	2.0	Honeywell	2.2
Altria Group	4.5	IBM	1.0
American Express	0.9	Intel	2.0
AT&T	4.7	Johnson & Johnson	2.3
Boeing	1.6	JPMorgan Chase	3.3
Caterpillar	1.3	McDonald's	1.9
Citigroup	4.3	Merck	4.3
Coca-Cola	3.0	Microsoft	1.3
Disney	1.0	3M	2.5
DuPont	3.6	Pfizer	3.7
ExxonMobil	2.1	Procter & Gamble	1.9
General Electric	3.0	United Technologies	1.5
General Motors	5.2	Verizon	4.8
Hewlett-Packard	0.9	Wal-Mart Stores	1.3

putting, poor short game, and poor strategic decisions. The data obtained follow (*Golf Magazine,* February 2002):

CD file Golf

Mental approach	Mental approach	Short game	Short game	Short game
Practice	Accuracy	Mental approach	Accuracy	Putting
Power	Approach shots	Accuracy	Short game	Putting
Accuracy	Mental approach	Mental approach	Accuracy	Power
Accuracy	Accuracy	Short game	Power	Short game
Accuracy	Putting	Mental approach	Strategic decisions	Accuracy
Short game	Power	Mental approach	Approach shots	Short game
Practice	Practice	Mental approach	Power	Power
Mental approach	Short game	Mental approach	Short game	Strategic decisions
Accuracy	Short game	Accuracy	Mental approach	Short game
Mental approach	Putting	Mental approach	Mental approach	Putting
Practice	Putting	Practice	Short game	Putting
Power	Mental approach	Short game	Practice	Strategic decisions
Accuracy	Short game	Accuracy	Practice	Putting
Accuracy	Short game	Accuracy	Short game	Putting
Accuracy	Approach shots	Short game	Mental approach	Practice
Short game	Short game	Strategic decisions	Short game	Short game
Practice	Practice	Short game	Practice	Strategic decisions
Mental approach	Strategic decisions	Strategic decisions	Power	Short game
Accuracy	Practice	Practice	Practice	Accuracy

a. Develop a frequency and percent frequency distribution.
b. Which four critical areas most often prevent golfers from reaching their potential?

42. Ninety-four shadow stocks were reported by the American Association of Individual Investors. The term *shadow* indicates stocks for small to medium-sized firms not followed closely by the major brokerage houses. Information on where the stock was traded—New York Stock Exchange (NYSE), American Stock Exchange (AMEX), and over-the-counter (OTC)—the earnings per share, and the price/earnings ratio was provided for the following sample of 20 shadow stocks.

CD file Shadow

Stock	Exchange	Earnings per Share ($)	Price/Earnings Ratio
Chemi-Trol	OTC	.39	27.30
Candie's	OTC	.07	36.20
TST/Impreso	OTC	.65	12.70
Unimed Pharm.	OTC	.12	59.30
Skyline Chili	AMEX	.34	19.30
Cyanotech	OTC	.22	29.30
Catalina Light.	NYSE	.15	33.20
DDL Elect.	NYSE	.10	10.20
Euphonix	OTC	.09	49.70
Mesa Labs	OTC	.37	14.40
RCM Tech.	OTC	.47	18.60
Anuhco	AMEX	.70	11.40
Hello Direct	OTC	.23	21.10
Hilite Industries	OTC	.61	7.80
Alpha Tech.	OTC	.11	34.60
Wegener Group	OTC	.16	24.50
U.S. Home & Garden	OTC	.24	8.70
Chalone Wine	OTC	.27	44.40
Eng. Support Sys.	OTC	.89	16.70
Int. Remote Imaging	AMEX	.86	4.70

a. Provide frequency and relative frequency distributions for the exchange data. Where are most shadow stocks listed?

b. Provide frequency and relative frequency distributions for the earnings per share and price/earnings ratio data. Use classes of 0.00–0.19, 0.20–0.39, and so on for the earnings per share data and classes of 0.0–9.9, 10.0–19.9, and so on for the price/earnings ratio data. What observations and comments can you make about the shadow stocks?

43. Approximately 1.5 million high school students take the Scholastic Aptitude Test (SAT) each year and nearly 80% of the college and universities without open admissions policies use SAT scores in making admission decisions (College Board, March 2006). A sample of SAT scores for the combined math and verbal portions of the test are as follows:

CD file SATScores

1025	1042	1195	880	945
1102	845	1095	936	790
1097	913	1245	1040	998
998	940	1043	1048	1130
1017	1140	1030	1171	1035

a. Show a frequency distribution and histogram for the SAT scores. Begin the first class with an SAT score of 750 and use a class width of 100.
b. Comment on the shape of the distribution.
c. What other observations can be made about SAT scores based on the tabular and graphical summaries?

44. *Drug Store News* (September 2002) provided data on annual pharmacy sales for the leading pharmacy retailers in the United States. The following data are annual sales in millions.

Retailer	Sales	Retailer	Sales
Ahold USA	$ 1700	Medicine Shoppe	$ 1757
CVS	12700	Rite-Aid	8637
Eckerd	7739	Safeway	2150
Kmart	1863	Walgreens	11660
Kroger	3400	Wal-Mart	7250

a. Show a stem-and-leaf display.
b. Identify the annual sales levels for the smallest, medium, and largest drug retailers.
c. What are the two largest drug retailers?

45. Data from the U.S. Census Bureau provides the population by state in millions of people (*The World Almanac,* 2006).

CD file Population

State	Population	State	Population	State	Population
Alabama	4.5	Louisiana	4.5	Ohio	11.5
Alaska	0.7	Maine	1.3	Oklahoma	3.5
Arizona	5.7	Maryland	5.6	Oregon	3.6
Arkansas	2.8	Massachusetts	6.4	Pennsylvania	12.4
California	35.9	Michigan	10.1	Rhode Island	1.1
Colorado	4.6	Minnesota	5.1	South Carolina	4.2
Connecticut	3.5	Mississippi	2.9	South Dakota	0.8
Delaware	0.8	Missouri	5.8	Tennessee	5.9
Florida	17.4	Montana	0.9	Texas	22.5
Georgia	8.8	Nebraska	1.7	Utah	2.4
Hawaii	1.3	Nevada	2.3	Vermont	0.6
Idaho	1.4	New Hampshire	1.3	Virginia	7.5
Illinois	12.7	New Jersey	8.7	Washington	6.2
Indiana	6.2	New Mexico	1.9	West Virginia	1.8
Iowa	3.0	New York	19.2	Wisconsin	5.5
Kansas	2.7	North Carolina	8.5	Wyoming	0.5
Kentucky	4.1	North Dakota	0.6		

a. Develop a frequency distribution, a percent frequency distribution, and a histogram. Use a class width of 2.5 million.
b. Discuss the skewness in the distribution.
c. What observations can you make about the population of the 50 states?

46. Refer to the data set for high and low temperatures for 20 cities in exercise 47.
a. Develop a scatter diagram to show the relationship between the two variables, high temperature and low temperature.
b. Comment on the relationship between high and low temperatures.

47. The daily high and low temperatures for 20 cities follow (*USA Today,* March 3, 2006).

CD file CityTemp

City	High	Low	City	High	Low
Albuquerque	66	39	Los Angeles	60	46
Atlanta	61	35	Miami	84	65
Baltimore	42	26	Minneapolis	30	11
Charlotte	60	29	New Orleans	68	50
Cincinnati	41	21	Oklahoma City	62	40
Dallas	62	47	Phoenix	77	50
Denver	60	31	Portland	54	38
Houston	70	54	St. Louis	45	27
Indianapolis	42	22	San Francisco	55	43
Las Vegas	65	43	Seattle	52	36

a. Prepare a stem-and-leaf display of the high temperatures.
b. Prepare a stem-and-leaf display of the low temperatures.
c. Compare the two stem-and-leaf displays and make comments about the difference between the high and low temperatures.
d. Provide a frequency distribution for both high and low temperatures.

48. Do larger companies generate more revenue? The following data show the number of employees and annual revenue for a sample of 20 *Fortune* 1000 companies (*Fortune,* April 17, 2000).

CD file RevEmps

Company	Employees	Revenue ($ millions)	Company	Employees	Revenue ($ millions)
Sprint	77,600	19,930	American Financial	9,400	3,334
Chase Manhattan	74,801	33,710	Fluor	53,561	12,417
Computer Sciences	50,000	7,660	Phillips Petroleum	15,900	13,852
Wells Fargo	89,355	21,795	Cardinal Health	36,000	25,034
Sunbeam	12,200	2,398	Borders Group	23,500	2,999
CBS	29,000	7,510	MCI Worldcom	77,000	37,120
Time Warner	69,722	27,333	Consolidated Edison	14,269	7,491
Steelcase	16,200	2,743	IBP	45,000	14,075
Georgia-Pacific	57,000	17,796	Super Value	50,000	17,421
Toro	1,275	4,673	H&R Block	4,200	1,669

a. Prepare a scatter diagram to show the relationship between the variables Revenue and Employees.
b. Comment on any relationship between the variables.

49. A study of job satisfaction was conducted for four occupations. Job satisfaction was measured using an 18-item questionnaire with each question receiving a response score of 1 to 5 with higher scores indicating greater satisfaction. The sum of the 18 scores provides the job satisfaction score for each individual in the sample. The data are as follow.

CD file OccupSat

Occupation	Satisfaction Score	Occupation	Satisfaction Score	Occupation	Satisfaction Score
Lawyer	42	Physical Therapist	78	Systems Analyst	60
Physical Therapist	86	Systems Analyst	44	Physical Therapist	59
Lawyer	42	Systems Analyst	71	Cabinetmaker	78
Systems Analyst	55	Lawyer	50	Physical Therapist	60
Lawyer	38	Lawyer	48	Physical Therapist	50
Cabinetmaker	79	Cabinetmaker	69	Cabinetmaker	79
Lawyer	44	Physical Therapist	80	Systems Analyst	62
Systems Analyst	41	Systems Analyst	64	Lawyer	45
Physical Therapist	55	Physical Therapist	55	Cabinetmaker	84
Systems Analyst	66	Cabinetmaker	64	Physical Therapist	62
Lawyer	53	Cabinetmaker	59	Systems Analyst	73
Cabinetmaker	65	Cabinetmaker	54	Cabinetmaker	60
Lawyer	74	Systems Analyst	76	Lawyer	64
Physical Therapist	52				

a. Provide a crosstabulation of occupation and job satisfaction score.
b. Compute the row percentages for your crosstabulation in part (a).
c. What observations can you make concerning the level of job satisfaction for these occupations?

50. Table 2.17 contains a portion of the data on the file named Fortune on the CD that accompanies the text. It provides data on stockholders' equity, market value, and profits for a sample of 50 *Fortune* 500 companies.

TABLE 2.17 DATA FOR A SAMPLE OF 50 *FORTUNE* 500 COMPANIES

CD file Fortune

Company	Stockholders' Equity ($1000s)	Market Value ($1000s)	Profit ($1000s)
AGCO	982.1	372.1	60.6
AMP	2698.0	12017.6	2.0
Apple Computer	1642.0	4605.0	309.0
Baxter International	2839.0	21743.0	315.0
Bergen Brunswick	629.1	2787.5	3.1
Best Buy	557.7	10376.5	94.5
Charles Schwab	1429.0	35340.6	348.5
.	.	.	.
.	.	.	.
.	.	.	.
Walgreen	2849.0	30324.7	511.0
Westvaco	2246.4	2225.6	132.0
Whirlpool	2001.0	3729.4	325.0
Xerox	5544.0	35603.7	395.0

a. Prepare a crosstabulation for the variables Stockholders' Equity and Profit. Use classes of 0–200, 200–400, . . . , 1000–1200 for Profit, and classes of 0–1200, 1200–2400, . . . , 4800–6000 for Stockholders' Equity.
b. Compute the row percentages for your crosstabulation in part (a).
c. What relationship, if any, do you notice between Profit and Stockholders' Equity?

51. A survey of commercial buildings served by the Cinergy-Cincinnati Gas & Electric Company asked what main heating fuel was used and what year the building was constructed. A partial crosstabulation of the findings follows.

Year Constructed	Fuel Type				
	Electricity	Natural Gas	Oil	Propane	Other
1973 or before	40	183	12	5	7
1974–1979	24	26	2	2	0
1980–1986	37	38	1	0	6
1987–1991	48	70	2	0	1

a. Complete the crosstabulation by showing the row totals and column totals.
b. Show the frequency distributions for year constructed and for fuel type.
c. Prepare a crosstabulation showing column percentages.
d. Prepare a crosstabulation showing row percentages.
e. Comment on the relationship between year constructed and fuel type.

52. Refer to the data set in Table 2.17.
a. Prepare a scatter diagram to show the relationship between the variables Profit and Stockholders' Equity.
b. Comment on any relationship between the variables.

53. Refer to the data set in Table 2.17.
a. Prepare a scatter diagram to show the relationship between the variables Market Value and Stockholders' Equity.
b. Comment on any relationship between the variables.

54. Refer to the data set in Table 2.17.
a. Prepare a crosstabulation for the variables Market Value and Profit.
b. Compute the row percentages for your crosstabulation in part (a).
c. Comment on any relationship between the variables.

Case Problem 1 Pelican Stores

Pelican Stores, a division of National Clothing, is a chain of women's apparel stores operating throughout the country. The chain recently ran a promotion in which discount coupons were sent to customers of other National Clothing stores. Data collected for a sample of 100 in-store credit card transactions at Pelican Stores during one day while the promotion was running are contained in the file named PelicanStores. Table 2.18 shows a portion of the data set. The Proprietary Card method of payment refers to charges made using a National Clothing charge card. Customers who made a purchase using a discount coupon are referred to as promotional customers and customers who made a purchase but did not use a discount coupon are referred to as regular customers. Because the promotional coupons were not sent to regular Pelican Stores customers, management considers the sales made to people presenting the promotional coupons as sales it would not otherwise make. Of course, Pelican also hopes that the promotional customers will continue to shop at its stores.

CD file
PelicanStores

TABLE 2.18 DATA FOR A SAMPLE OF 100 CREDIT CARD PURCHASES AT PELICAN STORES

Customer	Type of Customer	Items	Net Sales	Method of Payment	Gender	Marital Status	Age
1	Regular	1	39.50	Discover	Male	Married	32
2	Promotional	1	102.40	Proprietary Card	Female	Married	36
3	Regular	1	22.50	Proprietary Card	Female	Married	32
4	Promotional	5	100.40	Proprietary Card	Female	Married	28
5	Regular	2	54.00	MasterCard	Female	Married	34
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
96	Regular	1	39.50	MasterCard	Female	Married	44
97	Promotional	9	253.00	Proprietary Card	Female	Married	30
98	Promotional	10	287.59	Proprietary Card	Female	Married	52
99	Promotional	2	47.60	Proprietary Card	Female	Married	30
100	Promotional	1	28.44	Proprietary Card	Female	Married	44

Most of the variables shown in Table 2.18 are self-explanatory, but two of the variables require some clarification.

Items	The total number of items purchased
Net Sales	The total amount ($) charged to the credit card

Pelican's management would like to use this sample data to learn about its customer base and to evaluate the promotion involving discount coupons.

Managerial Report

Use the tabular and graphical methods of descriptive statistics to help management develop a customer profile and to evaluate the promotional campaign. At a minimum, your report should include the following:

1. Percent frequency distribution for key variables.
2. A bar graph or pie chart showing the number of customer purchases attributable to the method of payment.
3. A crosstabulation of type of customer (regular or promotional) versus net sales. Comment on any similarities or differences present.
4. A scatter diagram to explore the relationship between net sales and customer age.

Case Problem 2 Motion Picture Industry

The motion picture industry is a competitive business. More than 50 studios produce a total of 300 to 400 new motion pictures each year, and the financial success of each motion picture varies considerably. The opening weekend gross sales ($ millions), the total gross sales ($ millions), the number of theaters the movie was shown in, and the number of weeks the motion picture was in the top 60 for gross sales are common variables used to measure the success of a motion picture. Data collected for a sample of 100 motion pictures produced

TABLE 2.19 PERFORMANCE DATA FOR 10 MOTION PICTURES

CD file Movies

Motion Picture	Opening Weekend Gross Sales ($ millions)	Total Gross Sales ($ millions)	Number of Theaters	Weeks in Top 60
Coach Carter	29.17	67.25	2574	16
Ladies in Lavender	0.15	6.65	119	22
Batman Begins	48.75	205.28	3858	18
Unleashed	10.90	24.47	1962	8
Pretty Persuasion	0.06	0.23	24	4
Fever Pitch	12.40	42.01	3275	14
Harry Potter and the Goblet of Fire	102.69	287.18	3858	13
Monster-in-Law	23.11	82.89	3424	16
White Noise	24.11	55.85	2279	7
Mr. and Mrs. Smith	50.34	186.22	3451	21

in 2005 are contained in the file named Movies. Table 2.19 shows the data for the first 10 motion pictures in this file.

Managerial Report

Use the tabular and graphical methods of descriptive statistics to learn how these variables contribute to the success of a motion picture. Include the following in your report.

1. Tabular and graphical summaries for each of the four variables along with a discussion of what each summary tells us about the motion picture industry.
2. A scatter diagram to explore the relationship between Total Gross Sales and Opening Weekend Gross Sales. Discuss.
3. A scatter diagram to explore the relationship between Total Gross Sales and Number of Theaters. Discuss.
4. A scatter diagram to explore the relationship between Total Gross Sales and Weeks in Top 60. Discuss.

Appendix 2.1 Using Minitab for Tabular and Graphical Presentations

Minitab offers extensive capabilities for constructing tabular and graphical summaries of data. In this appendix we show how Minitab can be used to construct several graphical summaries and the tabular summary of a crosstabulation. The graphical methods presented include the dot plot, the histogram, the stem-and-leaf display, and the scatter diagram.

Dot Plot

CD file Audit

We use the audit time data in Table 2.4 to demonstrate. The data are in column C1 of a Minitab worksheet. The following steps will generate a dot plot.

Step 1. Select the **Graph** menu and choose **Dotplot**
Step 2. Select **One Y, Simple** and click **OK**
Step 3. When the Dotplot-One Y, Simple dialog box appears:
Enter C1 in the **Graph Variables** box
Click **OK**

Histogram

CD file Audit

We show how to construct a histogram with frequencies on the vertical axis using the audit time data in Table 2.4. The data are in column C1 of a Minitab worksheet. The following steps will generate a histogram for audit times.

Step 1. Select the **Graph** menu
Step 2. Choose **Histogram**
Step 3. Select **Simple** and click **OK**
Step 4. When the Histogram-Simple dialog box appears:
Enter C1 in the **Graph Variables** box
Click **OK**
Step 5. When the Histogram appears:
Position the mouse pointer over any one of the bars
Double-click
Step 6. When the Edit Bars dialog box appears:
Click on the **Binning** tab
Select **Cutpoint** for Interval Type
Select **Midpoint/Cutpoint positions** for Interval Definition
Enter 10:35/5 in the **Midpoint/Cutpoint positions** box*
Click **OK**

Note that Minitab also provides the option of scaling the x-axis so that the numerical values appear at the midpoints of the histogram rectangles. If this option is desired, modify step 6 to include Select **Midpoint** for Interval Type and Enter 12:32/5 in the **Midpoint/Cutpoint positions** box. These steps provide the same histogram with the midpoints of the histogram rectangles labeled 12, 17, 22, 27, and 32.

Stem-and-Leaf Display

CD file ApTest

We use the aptitude test data in Table 2.8 to demonstrate the construction of a stem-and-leaf display. The data are in column C1 of a Minitab worksheet. The following steps will generate the stretched stem-and-leaf display shown in Section 2.3.

Step 1. Select the **Graph** menu
Step 2. Choose **Stem-and-Leaf**
Step 3. When the Stem-and-Leaf dialog box appears:
Enter C1 in the **Graph Variables** box
Click **OK**

Scatter Diagram

CD file Stereo

We use the stereo and sound equipment store data in Table 2.12 to demonstrate the construction of a scatter diagram. The weeks are numbered from 1 to 10 in column C1, the data for number of commercials are in column C2, and the data for sales are in column C3 of a Minitab worksheet. The following steps will generate the scatter diagram shown in Figure 2.7.

*The entry 10:35/5 indicates that 10 is the starting value for the histogram, 35 is the ending value for the histogram, and 5 is the class width.

Step 1. Select the **Graph** menu
Step 2. Choose **Scatterplot**
Step 3. Select **Simple** and click **OK**
Step 4. When the Scatterplot-Simple dialog box appears:
Enter C3 under **Y variables** and C2 under **X variables**
Click **OK**

Crosstabulation

CD file Restaurant

We use the data from Zagat's restaurant review, part of which is shown in Table 2.9, to demonstrate. The restaurants are numbered from 1 to 300 in column C1 of the Minitab worksheet. The quality ratings are in column C2, and the meal prices are in column C3.

Minitab can only create a crosstabulation for qualitative variables and meal price is a quantitative variable. So we need to first code the meal price data by specifying the class to which each meal price belongs. The following steps will code the meal price data to create four classes of meal price in column C4: \$10–19, \$20–29, \$30–39, and \$40–49.

Step 1. Select the **Data** menu
Step 2. Choose **Code**
Step 3. Choose **Numeric to Text**
Step 4. When the Code-Numeric to Text dialog box appears:
Enter C3 in the **Code data from columns** box
Enter C4 in the **Into columns** box
Enter 10:19 in the first **Original values** box and \$10-19 in the adjacent **New** box
Enter 20:29 in the second **Original values** box and \$20-29 in the adjacent **New** box
Enter 30:39 in the third **Original values** box and \$30-39 in the adjacent **New** box
Enter 40:49 in the fourth **Original values** box and \$40-49 in the adjacent **New** box
Click **OK**

For each meal price in column C3 the associated meal price category will now appear in column C4. We can now develop a crosstabulation for quality rating and the meal price categories by using the data in columns C2 and C4. The following steps will create a crosstabulation containing the same information as shown in Table 2.10.

Step 1. Select the **Stat** menu
Step 2. Choose **Tables**
Step 3. Choose **Cross Tabulation and Chi-Square**
Step 4. When the Cross Tabulation and Chi-Square dialog box appears:
Enter C2 in the **For rows** box and C4 in the **For columns** box
Select **Counts** under Display
Click **OK**

Appendix 2.2 Using Excel for Tabular and Graphical Presentations

The Microsoft® Excel appendixes throughout the text show how to use Excel 2007, the most recent version of Excel.

Excel offers extensive capabilities for constructing tabular and graphical summaries of data. In this appendix, we show how Excel can be used to construct a frequency distribution, bar graph, pie chart, histogram, scatter diagram, and crosstabulation. We will demonstrate two of Excel's most powerful tools for data analysis: creating charts and creating PivotTable Reports.

Frequency Distribution and Bar Graph for Qualitative Data

In this section we show how Excel can be used to construct a frequency distribution and a bar graph for qualitative data. We illustrate each using the data on soft drink purchases in Table 2.1.

Frequency distribution We begin by showing how the COUNTIF function can be used to construct a frequency distribution for the data in Table 2.1. Refer to Figure 2.10 as we describe the steps involved. The formula worksheet (showing the function used) is set in the background, and the value worksheet (showing the results obtained using the function) appears in the foreground.

CD file SoftDrink

The label "Brand Purchased" and the data for the 50 soft drink purchases are in cells A1:A51. We also entered the labels "Soft Drink" and "Frequency" in cells C1:D1. The five soft drink names are entered into cells C2:C6. Excel's COUNTIF function can now be used to count the number of times each soft drink appears in cells A2:A51. The following steps are used.

Step 1. Select cell D2
Step 2. Enter =COUNTIF(\$A\$2:\$A\$51,C2)
Step 3. Copy cell D2 to cells D3:D6

FIGURE 2.10 FREQUENCY DISTRIBUTION FOR SOFT DRINK PURCHASES CONSTRUCTED USING EXCEL'S COUNTIF FUNCTION

	A	B	C	D	E
1	**Brand Purchased**		**Soft Drink**	**Frequency**	
2	Coke Classic		Coke Classic	=COUNTIF(\$A\$2:\$A\$51,C2)	
3	Diet Coke		Diet Coke	=COUNTIF(\$A\$2:\$A\$51,C3)	
4	Pepsi		Dr. Pepper	=COUNTIF(\$A\$2:\$A\$51,C4)	
5	Diet Coke		Pepsi	=COUNTIF(\$A\$2:\$A\$51,C5)	
6	Coke Classic		Sprite	=COUNTIF(\$A\$2:\$A\$51,C6)	
7	Coke Classic				
8	Dr. Pepper				
9	Diet Coke				
10	Pepsi				
45	Pepsi				
46	Pepsi				
47	Pepsi				
48	Coke Classic				
49	Dr. Pepper				
50	Pepsi				
51	Sprite				
52					

Note: Rows 11–44 are hidden.

	A	B	C	D	E
1	**Brand Purchased**		**Soft Drink**	**Frequency**	
2	Coke Classic		Coke Classic	19	
3	Diet Coke		Diet Coke	8	
4	Pepsi		Dr. Pepper	5	
5	Diet Coke		Pepsi	13	
6	Coke Classic		Sprite	5	
7	Coke Classic				
8	Dr. Pepper				
9	Diet Coke				
10	Pepsi				
45	Pepsi				
46	Pepsi				
47	Pepsi				
48	Coke Classic				
49	Dr. Pepper				
50	Pepsi				
51	Sprite				
52					

The formula worksheet in Figure 2.10 shows the cell formulas inserted by applying these steps. The value worksheet shows the values computed by the cell formulas. This worksheet shows the same frequency distribution that we developed in Table 2.2.

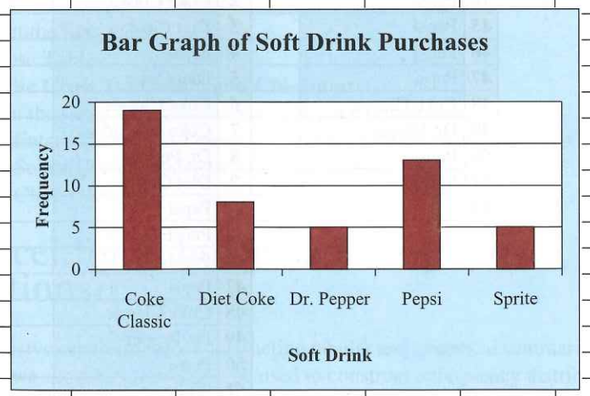
CD file SoftDrink

Bar graph Here we show how Excel's charting capability can be used to construct a bar graph for the soft drink data. Refer to the frequency distribution shown in the value worksheet of Figure 2.10. The bar chart that we are going to develop is an extension of this worksheet. The worksheet and the bar graph developed are shown in Figure 2.11. The steps are as follows:

Step 1. Select cells C2:D6
Step 2. Click the **Insert** tab on the Ribbon
Step 3. In the **Charts** group, click **Column**
Step 4. When the list of column chart subtypes appears:
Go to the **2-D Column** section
Click **Clustered Column** (the leftmost chart)
Step 5. In the **Chart Layouts** group, click the **More** button (the downward-pointing arrow with a line over it) to display all the options
Step 6. Choose **Layout 9**
Step 7. Select **Chart Title** and replace it with Bar Graph of Soft Drink Purchases
Step 8. Select the horizontal **Axis Title** and replace it with Soft Drink
Step 9. Select the vertical **Axis Title** and replace it with Frequency
Step 10. Right-click on the legend (**Series 1**)
Select **Delete**

FIGURE 2.11 BAR GRAPH OF SOFT DRINK PURCHASES CONSTRUCTED USING EXCEL

	A	B	C	D	E	F	G	H	I
1	**Brand Purchased**		**Soft Drink**	**Frequency**					
2	Coke Classic		Coke Classic	19					
3	Diet Coke		Diet Coke	8					
4	Pepsi		Dr. Pepper	5					
5	Diet Coke		Pepsi	13					
6	Coke Classic		Sprite	5					
7	Coke Classic								
8	Dr. Pepper								
9	Diet Coke								
10	Pepsi								
45	Pepsi								
46	Pepsi								
47	Pepsi								
48	Coke Classic								
49	Dr. Pepper								
50	Pepsi								
51	Sprite								
52									
53									
54									
55									
56									
57									



Step 11. Right-click the vertical axis
Select **Format Axis**
Step 12. When the Format Axis dialog box appears:
Go to the **Axis Options** section
Select **Fixed** for **Major Unit** and enter 5.0 in the corresponding box
Click **Close**

The resulting bar graph is shown in Figure 2.11.*

Excel can produce a pie chart for the soft drink data in a similar fashion. The major difference is that in step 3 we would click **Pie** in the Charts group. Several styles of pie charts are available.

Frequency Distribution and Histogram for Quantitative Data

In this section we show how Excel can be used to construct a frequency distribution and a histogram for quantitative data. We illustrate each using the audit time data shown in Table 2.4.

CD file Audit

Frequency distribution Excel's FREQUENCY function can be used to construct a frequency distribution for quantitative data. Refer to Figure 2.12 as we describe the steps involved. The formula worksheet is in the background, and the value worksheet is in the foreground. The label "Audit Time" is in cell A1 and the data for the 20 audits are in cells A2:A21. Using the procedures described in the text, we make the five classes 10–14, 15–19, 20–24, 25–29, and 30–34. The label "Audit Time" and the five classes are entered in cells C1:C6. The label "Upper Limit" and the five class upper limits are entered in cells D1:D6. We also entered the label "Frequency" in cell E1. Excel's FREQUENCY function will be used to show the class frequencies in cells E2:E6. The following steps describe how to develop a frequency distribution for the audit time data.

Step 1. Select cells E2:E6
Step 2. Type, but do not enter, the following formula:

=FREQUENCY(A2:A21,D2:D6)

You must hold down the Ctrl and Shift keys while pressing the Enter key to enter an array formula.

Step 3. Press CTRL + SHIFT + ENTER and the array formula will be entered into each of the cells E2:E6

The results are shown in Figure 2.12. The values displayed in the cells E2:E6 indicate frequencies for the corresponding classes. Referring to the FREQUENCY function, we see that the range of cells for the upper class limits (D2:D6) provides input to the function. These upper class limits, which Excel refers to as *bins,* tell Excel which frequency to put into the cells of the output range (E2:E6). For example, the frequency for the class with an upper limit, or bin, of 14 is placed in the first cell (E2), the frequency for the class with an upper limit, or bin, of 19 is placed in the second cell (E3), and so on.

Histogram To use Excel to construct a histogram for the audit time data, we begin with the frequency distribution as shown in Figure 2.12. The frequency distribution worksheet and the histogram output are shown in Figure 2.13. The following steps describe how to construct a histogram from a frequency distribution.

Step 1. Select cells C2:C6
Step 2. Press the Ctrl key and also select cells E2:E6

*The bar graph in Figure 2.11 can be resized. Resizing an Excel chart is not difficult. First, select the chart. Sizing handles will appear on the chart border. Click on the sizing handles and drag them to resize the figure to your preference.

FIGURE 2.12 FREQUENCY DISTRIBUTION FOR AUDIT TIME DATA CONSTRUCTED USING EXCEL'S FREQUENCY FUNCTION

	A	B	C	D	E
1	Audit Time		Audit Time	Upper Limit	Frequency
2	12		10-14	14	=FREQUENCY(A2:A21,D2:D6)
3	15		15-19	19	=FREQUENCY(A2:A21,D2:D6)
4	20		20-24	24	=FREQUENCY(A2:A21,D2:D6)
5	22		25-29	29	=FREQUENCY(A2:A21,D2:D6)
6	14		30-34	34	=FREQUENCY(A2:A21,D2:D6)
7	14				
8	15				
9	27				
10	21				
11	18				
12	19				
13	18				
14	22				
15	33				
16	16				
17	18				
18	17				
19	23				
20	28				
21	13				

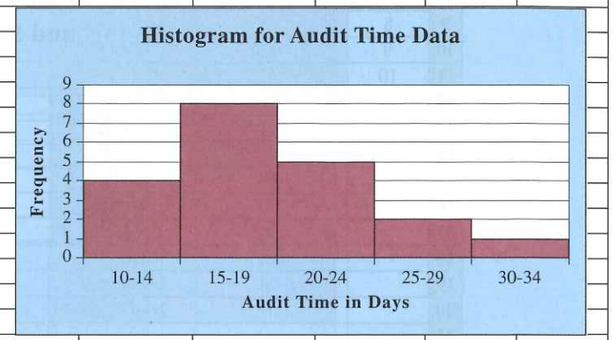
	A	B	C	D	E
1	Audit Times		Audit Times	Upper Limit	Frequency
2	12		10-14	14	4
3	15		15-19	19	8
4	20		20-24	24	5
5	22		25-29	29	2
6	14		30-34	34	1
7	14				
8	15				
9	27				
10	21				
11	18				
12	19				
13	18				
14	22				
15	33				
16	16				
17	18				
18	17				
19	23				
20	28				
21	13				

Step 3. Click the **Insert** tab on the Ribbon
Step 4. In the **Charts** group, click **Column**
Step 5. When the list of column chart subtypes appears:
Go to the **2-D Column** section
Click **Clustered Column** (the leftmost chart)
Step 6. In the **Chart Layouts** group, click the **More** button (the downward-pointing arrow with a line over it)
Step 7. Choose **Layout 8**
Step 8. Select **Chart Title** and replace it with Histogram for Audit Time Data
Step 9. Select the horizontal **Axis Title** and replace it with Audit Time in Days
Step 10. Select the vertical **Axis Title** and replace it with Frequency

Finally, an interesting aspect of the worksheet in Figure 2.13 is that Excel links the data in cells A2:A21 to the frequencies in cells E2:E6 and to the histogram. If an edit or revision of the data in cells A2:A21 occurs, the frequencies in cells E2:E6 and the histogram will be updated automatically to display a revised frequency distribution and histogram. Try one or two data edits to see how this automatic updating works.

FIGURE 2.13 HISTOGRAM FOR THE AUDIT TIME DATA CONSTRUCTED USING EXCEL

	A	B	C	D	E	F	G
1	Audit Time		Audit Time	Upper Limit	Frequency		
2	12		10-14	14	4		
3	15		15-19	19	8		
4	20		20-24	24	5		
5	22		25-29	29	2		
6	14		30-34	34	1		
7	14						
8	15						
9	27						
10	21						
11	18						
12	19						
13	18						
14	22						
15	33						
16	16						
17	18						
18	17						
19	23						
20	28						
21	13						
22							



Scatter Diagram

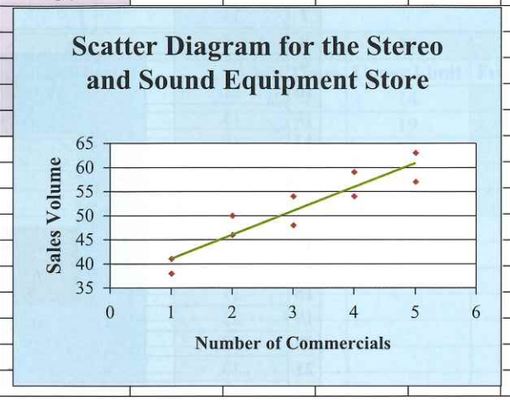
CD file Stereo

We use the stereo and sound equipment store data in Table 2.12 to demonstrate the use of Excel to construct a scatter diagram. Refer to Figure 2.14 as we describe the tasks involved. The value worksheet is set in the background, and the scatter diagram produced by Excel appears in the foreground. The following steps will produce the scatter diagram.

Step 1. Select cells B2:C11
Step 2. Click the **Insert** tab on the Ribbon
Step 3. In the **Charts** group, click **Scatter**
Step 4. When the list of scatter diagram subtypes appears:
Click **Scatter with only Markers** (the chart in the upper left corner)
Step 5. In the **Chart Layouts** group, click **Layout 1**
Step 6. Select **Chart Title** and replace it with Scatter Diagram for the Stereo and Sound Equipment Store
Step 7. Select the horizontal **Axis Title** and replace it with Number of Commercials
Step 8. Select the vertical **Axis Title** and replace it with Sales Volume
Step 9. Right-click the legend **Series 1**
Select **Delete**
Step 10. Right-click the vertical axis
Select **Format Axis**
Step 11. When the Format Axis dialog box appears:
Go to the **Axis Options** section
Select **Fixed** for **Minimum** and enter 35 in the corresponding box
Select **Fixed** for **Maximum** and enter 65 in the corresponding box
Select **Fixed** for **Major Unit** and enter 5 in the corresponding box
Click **Close**

FIGURE 2.14 SCATTER DIAGRAM FOR STEREO AND SOUND EQUIPMENT STORE USING EXCEL

	A	B	C	D	E	F	G	H
1	Week	No. of Commercials	Sales Volume					
2	1	2	50					
3	2	5	57					
4	3	1	41					
5	4	3	54					
6	5	4	54					
7	6	1						
8	7	5						
9	8	3						
10	9	4						
11	10	2						
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								



A trendline can be added to the scatter diagram as follows.

Step 1. Position the mouse pointer over any data point in the scatter diagram and right-click to display a list of options
Step 2. Choose **Add Trendline**
Step 3. When the **Format Trendline** dialog box appears:
Go to the **Trendline Options** section
Choose **Linear** in the **Trend/Regression type** section
Click **Close**

The worksheet in Figure 2.14 shows the scatter diagram with the trendline added.

PivotTable Report

Excel's PivotTable Report provides a valuable tool for managing data sets involving more than one variable. We will illustrate its use by showing how to develop a crosstabulation using the restaurant data in Figure 2.15. Labels are entered in row 1, and the data for each of the 300 restaurants are entered into cells A2:C301.

Creating the initial worksheet The following steps are needed to create a worksheet containing the initial PivotTable Report and PivotTable Field List.

Step 1. Click the **Insert** tab on the Ribbon
Step 2. In the **Tables** group, click the icon above PivotTable
Step 3. When the Create PivotTable dialog box appears:
Choose **Select a table or range**
Enter A1:C301 in the **Table/Range** box

FIGURE 2.15 EXCEL WORKSHEET CONTAINING RESTAURANT DATA

CD file
Restaurant

Note: Rows 12–291 are hidden.

	A	B	C	D
1	Restaurant	Quality Rating	Meal Price ($)	
2	1	Good	18	
3	2	Very Good	22	
4	3	Good	28	
5	4	Excellent	38	
6	5	Very Good	33	
7	6	Good	28	
8	7	Very Good	19	
9	8	Very Good	11	
10	9	Very Good	23	
11	10	Good	13	
292	291	Very Good	23	
293	292	Very Good	24	
294	293	Excellent	45	
295	294	Good	14	
296	295	Good	18	
297	296	Good	17	
298	297	Good	16	
299	298	Good	15	
300	299	Very Good	38	
301	300	Very Good	31	
302				

Select **New Worksheet**
Click **OK**

The resulting initial PivotTable Report and PivotTable Field List are shown in Figure 2.16.

Using the PivotTable Field List Each column in Figure 2.15 (Restaurant, Quality Rating, and Meal Price) is considered a field by Excel. The following steps show how to use Excel's PivotTable Field List to move the Quality Rating field to the row section, the Meal Price ($) field to the column section, and the Restaurant field to the values section of the PivotTable Report.

Step 1. In the **PivotTable Field List,** go to **Choose Fields to add to report:**
Drag the **Quality Rating** field to the **Row Labels** area
Drag the **Meal Price ($)** field to the **Column Labels** area
Drag the **Restaurant** field to the **Values** area
Step 2. Click **Sum of Restaurant** in the **Values** area
Select **Value Field Settings**
Step 3. When the Value Field Settings dialog appears:
Under **Summarize value field by,** choose **Count**
Click **OK**

Figure 2.17 shows the completed PivotTable Field List and a portion of the PivotTable Report.

Finalizing the PivotTable Report To complete the PivotTable Report we need to group the columns representing meal prices and place the row labels for quality rating in the proper order. The following steps accomplish these activities.

FIGURE 2.16 INITIAL PIVOTTABLE REPORT AND PIVOTTABLE FIELD LIST

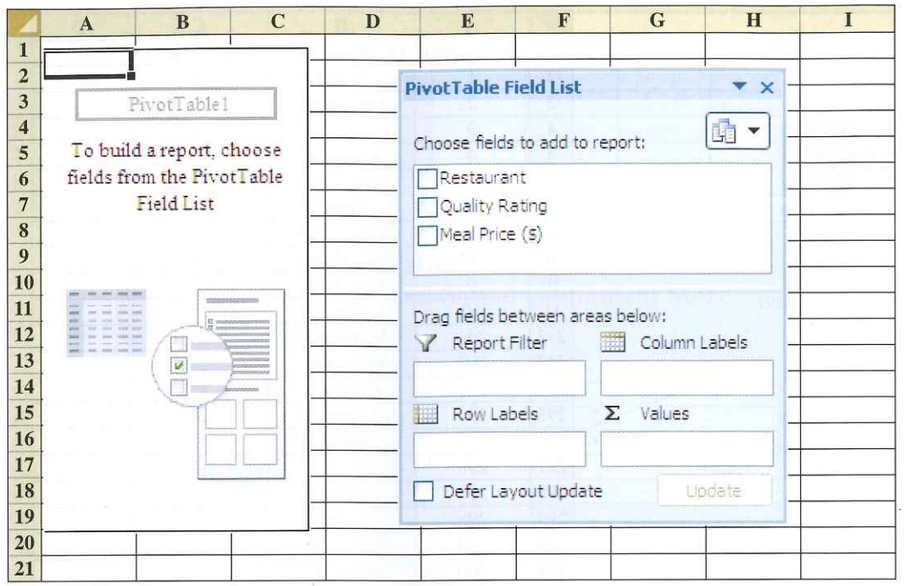


FIGURE 2.17 COMPLETED PIVOTTABLE FIELD LIST AND A PORTION OF PIVOTTABLE REPORT

	A	B	C	D	AL	AM	AN
1							
2							
3	Count of Restaurant	Column Labels					
4	Row Labels	10	11	12	47	48	Grand Total
5	Excellent				2	2	66
6	Good	6	4	3			84
7	Very Good	1	4	3		1	150
8	Grand Total	7	8	6	2	3	300
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							
20							

Note: Columns E-AK are hidden.

FIGURE 2.18 FINAL PIVOTTABLE REPORT

	A	B	C	D	E	F	G
1							
2							
3	Count of Restaurant	Column Labels					
4	Row Labels	10–19	20–29	30–39	40–49	Grand Total	
5	Good	42	40	2		84	
6	Very Good	34	64	46	6	150	
7	Excellent	2	14	28	22	66	
8	Grand Total	78	118	76	28	300	
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							
20							

Step 1. Right-click in cell B4 or in any other cell containing meal prices
Select **Group**

Step 2. When the Grouping dialog box appears:
Enter 10 in the **Starting at** box
Enter 49 in the **Ending at** box
Enter 10 in the **By** box
Click **OK**

Step 3. Right-click on **Excellent** in cell A5
Choose **Move**
Select **Move "Excellent" to End**

Step 4. Close the PivotTable Field List box

The final PivotTable Report is shown in Figure 2.18. Note that it provides the same information as the crosstabulation shown in Table 2.10.