CHAPTER 1

Data and Statistics

CONTENTS

STATISTICS IN PRACTICE: *BUSINESSWEEK*

1.1 APPLICATIONS IN BUSINESS AND ECONOMICS
Accounting
Finance
Marketing
Production
Economics

1.2 DATA
Elements, Variables, and Observations
Scales of Measurement
Categorical and Quantitative Data
Cross-Sectional and Time Series Data

1.3 DATA SOURCES
Existing Sources
Statistical Studies
Data Acquisition Errors

1.4 DESCRIPTIVE STATISTICS

1.5 STATISTICAL INFERENCE

1.6 COMPUTERS AND STATISTICAL ANALYSIS

1.7 DATA MINING

1.8 ETHICAL GUIDELINES FOR STATISTICAL PRACTICE

STATISTICS *in* PRACTICE

BUSINESSWEEK[*]
NEW YORK, NEW YORK

With a global circulation of more than 1 million, *BusinessWeek* is the most widely read business magazine in the world. More than 200 dedicated reporters and editors in 26 bureaus worldwide deliver a variety of articles of interest to the business and economic community. Along with feature articles on current topics, the magazine contains regular sections on International Business, Economic Analysis, Information Processing, and Science & Technology. Information in the feature articles and the regular sections helps readers stay abreast of current developments and assess the impact of those developments on business and economic conditions.

Most issues of *BusinessWeek* provide an in-depth report on a topic of current interest. Often, the in-depth reports contain statistical facts and summaries that help the reader understand the business and economic information. For example, the February 23, 2009 issue contained a feature article about the home foreclosure crisis, the March 17, 2009 issue included a discussion of when the stock market would begin to recover, and the May 4, 2009 issue had a special report on how to make pay cuts less painful. In addition, the weekly *BusinessWeek Investor* provides statistics about the state of the economy, including production indexes, stock prices, mutual funds, and interest rates.

BusinessWeek also uses statistics and statistical information in managing its own business. For example, an annual survey of subscribers helps the company learn about subscriber demographics, reading habits, likely purchases, lifestyles, and so on. *BusinessWeek* managers use statistical summaries from the survey to provide better services to subscribers and advertisers. One recent North American subscriber survey indicated that 90% of *BusinessWeek* subscribers use a personal computer at home and that 64% of *BusinessWeek* subscribers are involved with computer purchases at work. Such statistics alert *BusinessWeek* managers to subscriber interest in articles about new developments in computers. The results of the survey are also made available to potential advertisers. The high percentage of subscribers using personal computers at home and the high percentage of subscribers involved with computer purchases at work would be an incentive for a computer manufacturer to consider advertising in *BusinessWeek*.

In this chapter, we discuss the types of data available for statistical analysis and describe how the data are obtained. We introduce descriptive statistics and statistical inference as ways of converting data into meaningful and easily interpreted statistical information.

BusinessWeek uses statistical facts and summaries in many of its articles. © Terri Miller/E-Visual Communications, Inc.

[*] The authors are indebted to Charlene Trentham, Research Manager at *BusinessWeek*, for providing this Statistics in Practice.

Frequently, we see the following types of statements in newspapers and magazines:

- The National Association of Realtors reported that the median price paid by first-time home buyers is $165,000 (*The Wall Street Journal,* February 11, 2009).
- NCAA president Myles Brand reported that college athletes are earning degrees at record rates. Latest figures show that 79% of all men and women student-athletes graduate (Associated Press, October 15, 2008).
- The average one-way travel time to work is 25.3 minutes (U.S. Census Bureau, March 2009).

- A record high 11% of U.S. homes are vacant, a glut created by the housing boom and subsequent collapse (*USA Today,* February 13, 2009).
- The national average price for regular gasoline reached $4.00 per gallon for the first time in history (Cable News Network website*,* June 8, 2008).
- The New York Yankees have the highest salaries in major league baseball. The total payroll is $201,449,289 with a median salary of $5,000,000 (*USA Today Salary Data Base,* April 2009).
- The Dow Jones Industrial Average closed at 8721 (*The Wall Street Journal,* June 2, 2009).

The numerical facts in the preceding statements ($165,000, 79%, 25.3, 11%, $4.00, $201,449,289, $5,000,000 and 8721) are called statistics. In this usage, the term *statistics* refers to numerical facts such as averages, medians, percents, and index numbers that help us understand a variety of business and economic situations. However, as you will see, the field, or subject, of statistics involves much more than numerical facts. In a broader sense, **statistics** is defined as the art and science of collecting, analyzing, presenting, and interpreting data. Particularly in business and economics, the information provided by collecting, analyzing, presenting, and interpreting data gives managers and decision makers a better understanding of the business and economic environment and thus enables them to make more informed and better decisions. In this text, we emphasize the use of statistics for business and economic decision making.

Chapter 1 begins with some illustrations of the applications of statistics in business and economics. In Section 1.2 we define the term *data* and introduce the concept of a data set. This section also introduces key terms such as *variables* and *observations,* discusses the difference between quantitative and categorical data, and illustrates the uses of cross-sectional and time series data. Section 1.3 discusses how data can be obtained from existing sources or through survey and experimental studies designed to obtain new data. The important role that the Internet now plays in obtaining data is also highlighted. The uses of data in developing descriptive statistics and in making statistical inferences are described in Sections 1.4 and 1.5. The last three sections of Chapter 1 provide the role of the computer in statistical analysis, an introduction to the relative new field of data mining, and a discussion of ethical guidelines for statistical practice. A chapter-ending appendix includes an introduction to the add-in StatTools which can be used to extend the statistical options for users of Microsoft Excel.

1.1 Applications in Business and Economics

In today's global business and economic environment, anyone can access vast amounts of statistical information. The most successful managers and decision makers understand the information and know how to use it effectively. In this section, we provide examples that illustrate some of the uses of statistics in business and economics.

Accounting

Public accounting firms use statistical sampling procedures when conducting audits for their clients. For instance, suppose an accounting firm wants to determine whether the amount of accounts receivable shown on a client's balance sheet fairly represents the actual amount of accounts receivable. Usually the large number of individual accounts receivable makes reviewing and validating every account too time-consuming and expensive. As common practice in such situations, the audit staff selects a subset of the accounts called a sample. After reviewing the accuracy of the sampled accounts, the auditors draw a conclusion as to whether the accounts receivable amount shown on the client's balance sheet is acceptable.

Finance

Financial analysts use a variety of statistical information to guide their investment recommendations. In the case of stocks, the analysts review a variety of financial data including price/earnings ratios and dividend yields. By comparing the information for an individual stock with information about the stock market averages, a financial analyst can begin to draw a conclusion as to whether an individual stock is over- or underpriced. For example, *Barron's* (February 18, 2008) reported that the average dividend yield for the 30 stocks in the Dow Jones Industrial Average was 2.45%. Altria Group showed a dividend yield of 3.05%. In this case, the statistical information on dividend yield indicates a higher dividend yield for Altria Group than the average for the Dow Jones stocks. Therefore, a financial analyst might conclude that Altria Group was underpriced. This and other information about Altria Group would help the analyst make a buy, sell, or hold recommendation for the stock.

Marketing

Electronic scanners at retail checkout counters collect data for a variety of marketing research applications. For example, data suppliers such as ACNielsen and Information Resources, Inc., purchase point-of-sale scanner data from grocery stores, process the data, and then sell statistical summaries of the data to manufacturers. Manufacturers spend hundreds of thousands of dollars per product category to obtain this type of scanner data. Manufacturers also purchase data and statistical summaries on promotional activities such as special pricing and the use of in-store displays. Brand managers can review the scanner statistics and the promotional activity statistics to gain a better understanding of the relationship between promotional activities and sales. Such analyses often prove helpful in establishing future marketing strategies for the various products.

Production

Today's emphasis on quality makes quality control an important application of statistics in production. A variety of statistical quality control charts are used to monitor the output of a production process. In particular, an x-bar chart can be used to monitor the average output. Suppose, for example, that a machine fills containers with 12 ounces of a soft drink. Periodically, a production worker selects a sample of containers and computes the average number of ounces in the sample. This average, or x-bar value, is plotted on an x-bar chart. A plotted value above the chart's upper control limit indicates overfilling, and a plotted value below the chart's lower control limit indicates underfilling. The process is termed "in control" and allowed to continue as long as the plotted x-bar values fall between the chart's upper and lower control limits. Properly interpreted, an x-bar chart can help determine when adjustments are necessary to correct a production process.

Economics

Economists frequently provide forecasts about the future of the economy or some aspect of it. They use a variety of statistical information in making such forecasts. For instance, in forecasting inflation rates, economists use statistical information on such indicators as the Producer Price Index, the unemployment rate, and manufacturing capacity utilization. Often these statistical indicators are entered into computerized forecasting models that predict inflation rates.

Applications of statistics such as those described in this section are an integral part of this text. Such examples provide an overview of the breadth of statistical applications. To supplement these examples, practitioners in the fields of business and economics provided chapter-opening Statistics in Practice articles that introduce the material covered in each chapter. The Statistics in Practice applications show the importance of statistics in a wide variety of business and economic situations.

1.2 Data

Data are the facts and figures collected, analyzed, and summarized for presentation and interpretation. All the data collected in a particular study are referred to as the **data set** for the study. Table 1.1 shows a data set containing information for 25 mutual funds that are part of the *Morningstar Funds500* for 2008. Morningstar is a company that tracks over 7000 mutual funds and prepares in-depth analyses of 2000 of these. Their recommendations are followed closely by financial analysts and individual investors.

Elements, Variables, and Observations

Elements are the entities on which data are collected. For the data set in Table 1.1 each individual mutual fund is an element: the element names appear in the first column. With 25 mutual funds, the data set contains 25 elements.

A **variable** is a characteristic of interest for the elements. The data set in Table 1.1 includes the following five variables:

- *Fund Type:* The type of mutual fund, labeled DE (Domestic Equity), IE (International Equity), and FI (Fixed Income)
- *Net Asset Value ($):* The closing price per share on December 31, 2007

WEB file

Morningstar

Data sets such as Morningstar are available on the website for this text.

TABLE 1.1 DATA SET FOR 25 MUTUAL FUNDS

Fund Name	Fund Type	Net Asset Value ($)	5-Year Average Return (%)	Expense Ratio (%)	Morningstar Rank
American Century Intl. Disc	IE	14.37	30.53	1.41	3-Star
American Century Tax-Free Bond	FI	10.73	3.34	0.49	4-Star
American Century Ultra	DE	24.94	10.88	0.99	3-Star
Artisan Small Cap	DE	16.92	15.67	1.18	3-Star
Brown Cap Small	DE	35.73	15.85	1.20	4-Star
DFA U.S. Micro Cap	DE	13.47	17.23	0.53	3-Star
Fidelity Contrafund	DE	73.11	17.99	0.89	5-Star
Fidelity Overseas	IE	48.39	23.46	0.90	4-Star
Fidelity Sel Electronics	DE	45.60	13.50	0.89	3-Star
Fidelity Sh-Term Bond	FI	8.60	2.76	0.45	3-Star
Gabelli Asset AAA	DE	49.81	16.70	1.36	4-Star
Kalmar Gr Val Sm Cp	DE	15.30	15.31	1.32	3-Star
Marsico 21st Century	DE	17.44	15.16	1.31	5-Star
Mathews Pacific Tiger	IE	27.86	32.70	1.16	3-Star
Oakmark I	DE	40.37	9.51	1.05	2-Star
PIMCO Emerg Mkts Bd D	FI	10.68	13.57	1.25	3-Star
RS Value A	DE	26.27	23.68	1.36	4-Star
T. Rowe Price Latin Am.	IE	53.89	51.10	1.24	4-Star
T. Rowe Price Mid Val	DE	22.46	16.91	0.80	4-Star
Thornburg Value A	DE	37.53	15.46	1.27	4-Star
USAA Income	FI	12.10	4.31	0.62	3-Star
Vanguard Equity-Inc	DE	24.42	13.41	0.29	4-Star
Vanguard Sht-Tm TE	FI	15.68	2.37	0.16	3-Star
Vanguard Sm Cp Idx	DE	32.58	17.01	0.23	3-Star
Wasatch Sm Cp Growth	DE	35.41	13.98	1.19	4-Star

Source: Morningstar Funds500 (2008).

- *5-Year Average Return (%):* The average annual return for the fund over the past 5 years
- *Expense Ratio:* The percentage of assets deducted each fiscal year for fund expenses
- *Morningstar Rank:* The overall risk-adjusted star rating for each fund; Morningstar ranks go from a low of 1-Star to a high of 5-Stars

Measurements collected on each variable for every element in a study provide the data. The set of measurements obtained for a particular element is called an **observation**. Referring to Table 1.1 we see that the set of measurements for the first observation (American Century Intl. Disc) is IE, 14.37, 30.53, 1.41, and 3-Star. The set of measurements for the second observation (American Century Tax-Free Bond) is FI, 10.73, 3.34, 0.49, and 4-Star, and so on. A data set with 25 elements contains 25 observations.

Scales of Measurement

Data collection requires one of the following scales of measurement: nominal, ordinal, interval, or ratio. The scale of measurement determines the amount of information contained in the data and indicates the most appropriate data summarization and statistical analyses.

When the data for a variable consist of labels or names used to identify an attribute of the element, the scale of measurement is considered a **nominal scale**. For example, referring to the data in Table 1.1, we see that the scale of measurement for the Fund Type variable is nominal because DE, IE, and FI are labels used to identify the category or type of fund. In cases where the scale of measurement is nominal, a numeric code as well as nonnumeric labels may be used. For example, to facilitate data collection and to prepare the data for entry into a computer database, we might use a numeric code by letting 1 denote Domestic Equity, 2 denote International Equity, and 3 denote Fixed Income. In this case the numeric values 1, 2, and 3 identify the category of fund. The scale of measurement is nominal even though the data appear as numeric values.

The scale of measurement for a variable is called an **ordinal scale** if the data exhibit the properties of nominal data and the order or rank of the data is meaningful. For example, Eastside Automotive sends customers a questionnaire designed to obtain data on the quality of its automotive repair service. Each customer provides a repair service rating of excellent, good, or poor. Because the data obtained are the labels—excellent, good, or poor—the data have the properties of nominal data. In addition, the data can be ranked, or ordered, with respect to the service quality. Data recorded as excellent indicate the best service, followed by good and then poor. Thus, the scale of measurement is ordinal. As another example, note that the Morningstar Rank for the data in Table 1.1 is ordinal data. It provides a rank from 1 to 5-Stars based on Morningstar's assessment of the fund's risk-adjusted return. Ordinal data can also be provided using a numeric code, for example, your class rank in school.

The scale of measurement for a variable is an **interval scale** if the data have all the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure. Interval data are always numeric. Scholastic Aptitude Test (SAT) scores are an example of interval-scaled data. For example, three students with SAT math scores of 620, 550, and 470 can be ranked or ordered in terms of best performance to poorest performance. In addition, the differences between the scores are meaningful. For instance, student 1 scored $620 - 550 = 70$ points more than student 2, while student 2 scored $550 - 470 = 80$ points more than student 3.

The scale of measurement for a variable is a **ratio scale** if the data have all the properties of interval data and the ratio of two values is meaningful. Variables such as distance, height, weight, and time use the ratio scale of measurement. This scale requires that a zero value be included to indicate that nothing exists for the variable at the zero point.

For example, consider the cost of an automobile. A zero value for the cost would indicate that the automobile has no cost and is free. In addition, if we compare the cost of $30,000 for one automobile to the cost of $15,000 for a second automobile, the ratio property shows that the first automobile is $30,000/$15,000 = 2 times, or twice, the cost of the second automobile.

Categorical and Quantitative Data

Data can be classified as either categorical or quantitative. Data that can be grouped by specific categories are referred to as **categorical data**. Categorical data use either the nominal or ordinal scale of measurement. Data that use numeric values to indicate how much or how many are referred to as **quantitative data**. Quantitative data are obtained using either the interval or ratio scale of measurement.

The statistical method appropriate for summarizing data depends upon whether the data are categorical or quantitative.

A **categorical variable** is a variable with categorical data, and a **quantitative variable** is a variable with quantitative data. The statistical analysis appropriate for a particular variable depends upon whether the variable is categorical or quantitative. If the variable is categorical, the statistical analysis is limited. We can summarize categorical data by counting the number of observations in each category or by computing the proportion of the observations in each category. However, even when the categorical data are identified by a numerical code, arithmetic operations such as addition, subtraction, multiplication, and division do not provide meaningful results. Section 2.1 discusses ways for summarizing categorical data.

Arithmetic operations provide meaningful results for quantitative variables. For example, quantitative data may be added and then divided by the number of observations to compute the average value. This average is usually meaningful and easily interpreted. In general, more alternatives for statistical analysis are possible when data are quantitative. Section 2.2 and Chapter 3 provide ways of summarizing quantitative data.

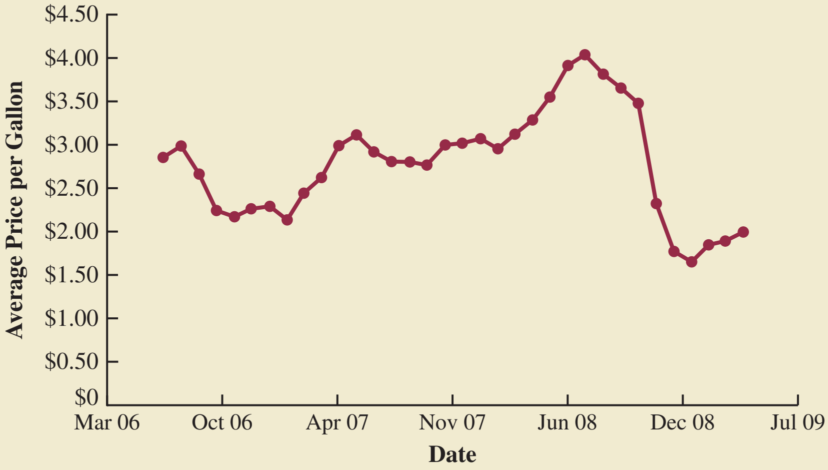
Cross-Sectional and Time Series Data

For purposes of statistical analysis, distinguishing between cross-sectional data and time series data is important. **Cross-sectional data** are data collected at the same or approximately the same point in time. The data in Table 1.1 are cross-sectional because they describe the five variables for the 25 mutual funds at the same point in time. **Time series data** are data collected over several time periods. For example, the time series in Figure 1.1 shows the U.S. average price per gallon of conventional regular gasoline between 2006 and 2009. Note that higher gasoline prices have tended to occur in the summer months, with the all-time-high average of $4.05 per gallon occurring in July 2008. By January 2009, gasoline prices had taken a steep decline to a three-year low of $1.65 per gallon.

Graphs of time series data are frequently found in business and economic publications. Such graphs help analysts understand what happened in the past, identify any trends over time, and project future levels for the time series. The graphs of time series data can take on a variety of forms, as shown in Figure 1.2. With a little study, these graphs are usually easy to understand and interpret.

For example, Panel (A) in Figure 1.2 is a graph that shows the Dow Jones Industrial Average Index from 1997 to 2009. In April 1997, the popular stock market index was near 7000. Over the next 10 years the index rose to over 14,000 in July 2007. However, notice the sharp decline in the time series after the all-time high in 2007. By March 2009, poor economic conditions had caused the Dow Jones Industrial Average Index to return to the 7000 level of 1997. This was a scary and discouraging period for investors. By June 2009, the index was showing a recovery by reaching 8700.

FIGURE 1.1 U.S. AVERAGE PRICE PER GALLON FOR CONVENTIONAL REGULAR GASOLINE



Source: Energy Information Administration, U.S. Department of Energy, May 2009.

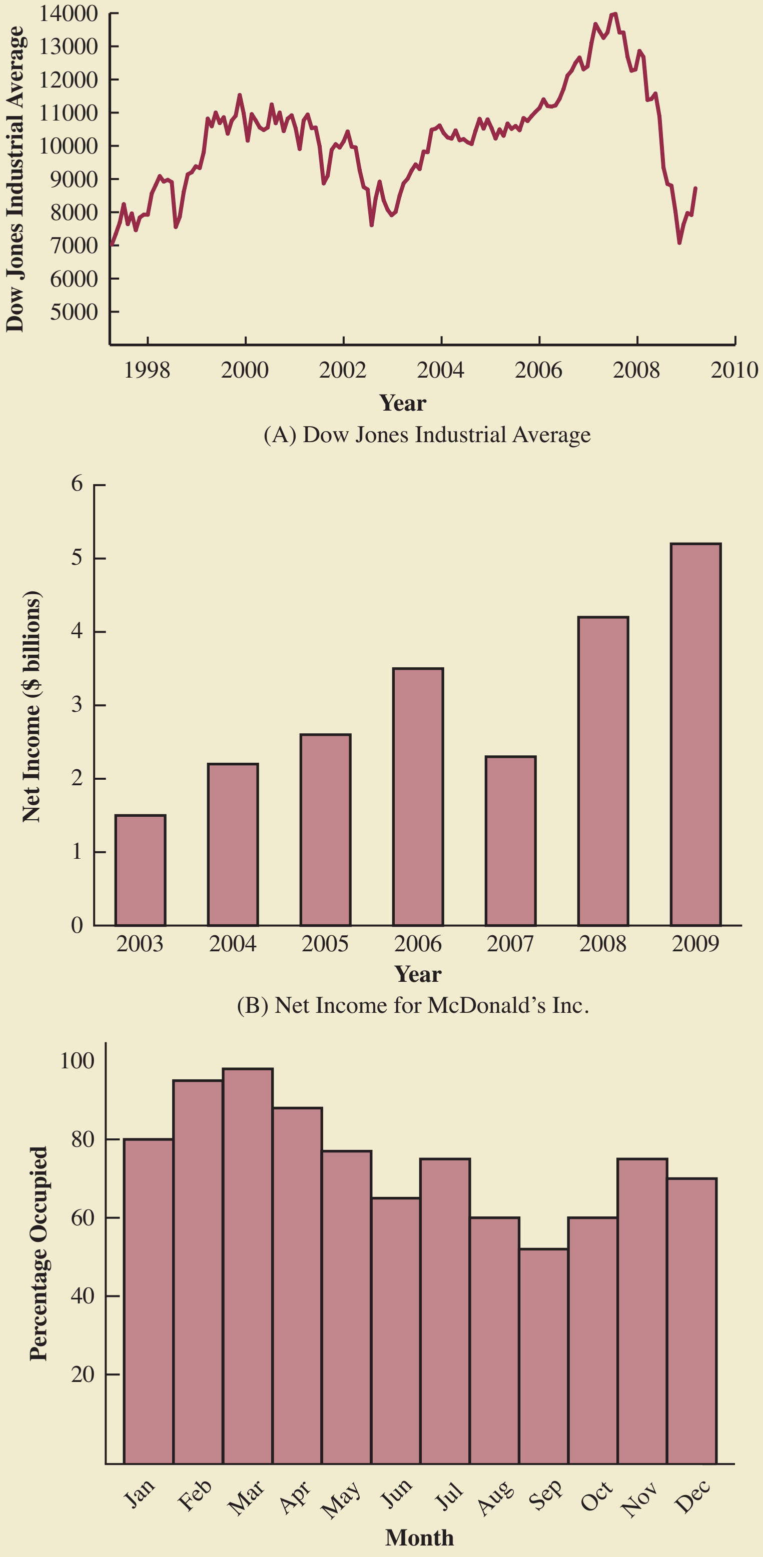
The graph in Panel (B) shows the net income of McDonald's Inc. from 2003 to 2009. The declining economic conditions in 2008 and 2009 were actually beneficial to McDonald's as the company's net income rose to an all-time high. The growth in McDonald's net income showed that the company was thriving during the economic downturn as people were cutting back on the more expensive sit-down restaurants and seeking less-expensive alternatives offered by McDonald's.

Panel (C) shows the time series for the occupancy rate of hotels in South Florida over a one-year period. The highest occupancy rates, 95% and 98%, occur during the months of February and March when the climate of South Florida is attractive to tourists. In fact, January to April of each year is typically the high-occupancy season for South Florida hotels. On the other hand, note the low occupancy rates during the months of August to October, with the lowest occupancy rate of 50% occurring in September. High temperatures and the hurricane season are the primary reasons for the drop in hotel occupancy during this period.

NOTES AND COMMENTS

1. An observation is the set of measurements obtained for each element in a data set. Hence, the number of observations is always the same as the number of elements. The number of measurements obtained for each element equals the number of variables. Hence, the total number of data items can be determined by multiplying the number of observations by the number of variables.
2. Quantitative data may be discrete or continuous. Quantitative data that measure how many (e.g., number of calls received in 5 minutes) are discrete. Quantitative data that measure how much (e.g., weight or time) are continuous because no separation occurs between the possible data values.

FIGURE 1.2 A VARIETY OF GRAPHS OF TIME SERIES DATA



(A) Dow Jones Industrial Average

(B) Net Income for McDonald's Inc.

(C) Occupancy Rate of South Florida Hotels

1.3 Data Sources

Data can be obtained from existing sources or from surveys and experimental studies designed to collect new data.

Existing Sources

In some cases, data needed for a particular application already exist. Companies maintain a variety of databases about their employees, customers, and business operations. Data on employee salaries, ages, and years of experience can usually be obtained from internal personnel records. Other internal records contain data on sales, advertising expenditures, distribution costs, inventory levels, and production quantities. Most companies also maintain detailed data about their customers. Table 1.2 shows some of the data commonly available from internal company records.

Organizations that specialize in collecting and maintaining data make available substantial amounts of business and economic data. Companies access these external data sources through leasing arrangements or by purchase. Dun & Bradstreet, Bloomberg, and Dow Jones & Company are three firms that provide extensive business database services to clients. ACNielsen and Information Resources, Inc., built successful businesses collecting and processing data that they sell to advertisers and product manufacturers.

Data are also available from a variety of industry associations and special interest organizations. The Travel Industry Association of America maintains travel-related information such as the number of tourists and travel expenditures by states. Such data would be of interest to firms and individuals in the travel industry. The Graduate Management Admission Council maintains data on test scores, student characteristics, and graduate management education programs. Most of the data from these types of sources are available to qualified users at a modest cost.

The Internet continues to grow as an important source of data and statistical information. Almost all companies maintain websites that provide general information about the company as well as data on sales, number of employees, number of products, product prices, and product specifications. In addition, a number of companies now specialize in making information available over the Internet. As a result, one can obtain access to stock quotes, meal prices at restaurants, salary data, and an almost infinite variety of information.

Government agencies are another important source of existing data. For instance, the U.S. Department of Labor maintains considerable data on employment rates, wage rates, size of the labor force, and union membership. Table 1.3 lists selected governmental agencies

TABLE 1.2 EXAMPLES OF DATA AVAILABLE FROM INTERNAL COMPANY RECORDS

Source	Some of the Data Typically Available
Employee records	Name, address, social security number, salary, number of vacation days, number of sick days, and bonus
Production records	Part or product number, quantity produced, direct labor cost, and materials cost
Inventory records	Part or product number, number of units on hand, reorder level, economic order quantity, and discount schedule
Sales records	Product number, sales volume, sales volume by region, and sales volume by customer type
Credit records	Customer name, address, phone number, credit limit, and accounts receivable balance
Customer profile	Age, gender, income level, household size, address, and preferences

TABLE 1.3 EXAMPLES OF DATA AVAILABLE FROM SELECTED GOVERNMENT AGENCIES

Government Agency	Some of the Data Available
Census Bureau	Population data, number of households, and household income
Federal Reserve Board	Data on the money supply, installment credit, exchange rates, and discount rates
Office of Management and Budget	Data on revenue, expenditures, and debt of the federal government
Department of Commerce	Data on business activity, value of shipments by industry, level of profits by industry, and growing and declining industries
Bureau of Labor Statistics	Consumer spending, hourly earnings, unemployment rate, safety records, and international statistics

and some of the data they provide. Most government agencies that collect and process data also make the results available through a website. Figure 1.3 shows the homepage for the U.S. Census Bureau website.

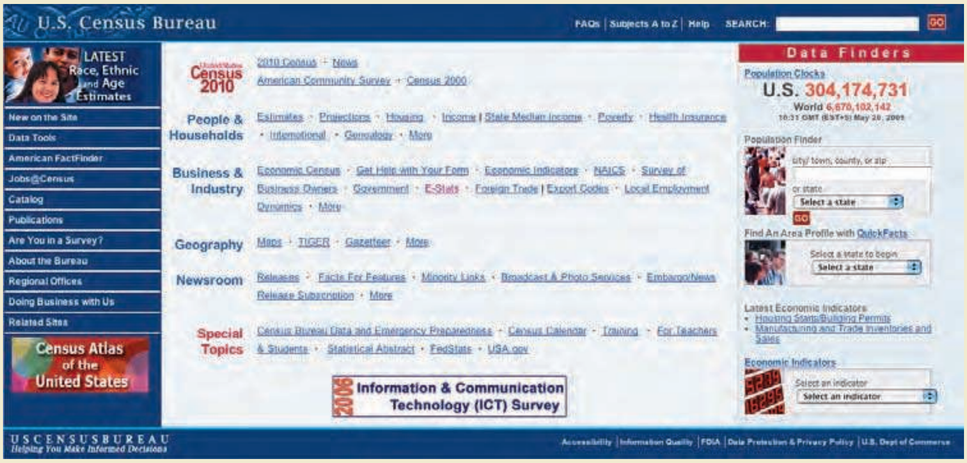
Statistical Studies

Sometimes the data needed for a particular application are not available through existing sources. In such cases, the data can often be obtained by conducting a statistical study. Statistical studies can be classified as either *experimental* or *observational.*

In an experimental study, a variable of interest is first identified. Then one or more other variables are identified and controlled so that data can be obtained about how they influence the variable of interest. For example, a pharmaceutical firm might be interested in conducting an experiment to learn about how a new drug affects blood pressure. Blood pressure is the variable of interest in the study. The dosage level of the new drug is another variable that is hoped to have a causal effect on blood pressure. To obtain data about the effect of the

The largest experimental statistical study ever conducted is believed to be the 1954 Public Health Service experiment for the Salk polio vaccine. Nearly 2 million children in grades 1, 2, and 3 were selected from throughout the United States.

FIGURE 1.3 U.S. CENSUS BUREAU HOMEPAGE


U.S. Census Bureau
FAQs | Subjects A to Z | Help SEARCH: GO
LATEST Race, Ethnic and Age Estimates
New on the Site
Data Tools
American FactFinder
Jobs@Census
Catalog
Publications
Are You in a Survey?
About the Bureau
Regional Offices
Doing Business with Us
Related Sites
Census Atlas of the United States
Census 2010
2010 Census · News
American Community Survey · Census 2000
People & Households
Estimates · Projections · Housing · Income | State Median Income · Poverty · Health Insurance · International · Genealogy · More
Business & Industry
Economic Census · Get Help with Your Form · Economic Indicators · NAICS · Survey of Business Owners · Government · E-Stats · Foreign Trade | Export Codes · Local Employment Dynamics · More
Geography
Maps · TIGER · Gazetteer · More
Newsroom
Releases · Facts For Features · Minority Links · Broadcast & Photo Services · Embargo/News Release Subscription · More
Special Topics
Census Bureau Data and Emergency Preparedness · Census Calendar · Training · For Teachers & Students · Statistical Abstract · FedStats · USA.gov
2006 Information & Communication Technology (ICT) Survey
Data Finders
Population Clocks
U.S. 304,174,731
World 6,670,102,142
16:31 GMT (EST+5) May 20, 2008
Population Finder
city/ town, county, or zip
or state
Select a state
GO
Find An Area Profile with QuickFacts
Select a state to begin
Select a state
Latest Economic Indicators
Housing Starts/Building Permits
Manufacturing and Trade Inventories and Sales
Economic Indicators
Select an indicator
Select an indicator
USCENSUSBUREAU
Helping You Make Informed Decisions
Accessibility | Information Quality | FOIA | Data Protection & Privacy Policy | U.S. Dept of Commerce

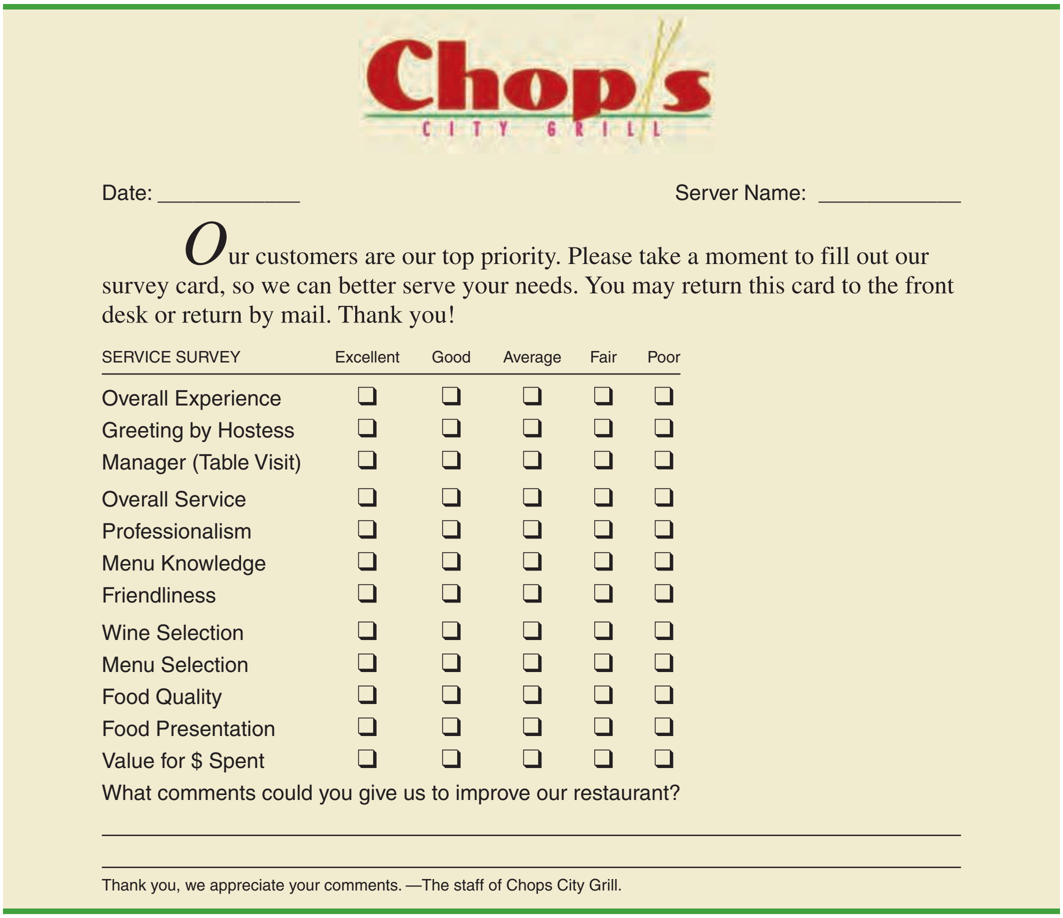
new drug, researchers select a sample of individuals. The dosage level of the new drug is controlled, as different groups of individuals are given different dosage levels. Before and after data on blood pressure are collected for each group. Statistical analysis of the experimental data can help determine how the new drug affects blood pressure.

Studies of smokers and nonsmokers are observational studies because researchers do not determine or control who will smoke and who will not smoke.

Nonexperimental, or observational, statistical studies make no attempt to control the variables of interest. A survey is perhaps the most common type of observational study. For instance, in a personal interview survey, research questions are first identified. Then a questionnaire is designed and administered to a sample of individuals. Some restaurants use observational studies to obtain data about customer opinions on the quality of food, quality of service, atmosphere, and so on. A customer opinion questionnaire used by Chops City Grill in Naples, Florida, is shown in Figure 1.4. Note that the customers who fill out the questionnaire are asked to provide ratings for 12 variables, including overall experience, greeting by hostess, manager (table visit), overall service, and so on. The response categories of excellent, good, average, fair, and poor provide categorical data that enable Chops City Grill management to maintain high standards for the restaurant's food and service.

Anyone wanting to use data and statistical analysis as aids to decision making must be aware of the time and cost required to obtain the data. The use of existing data sources is desirable when data must be obtained in a relatively short period of time. If important data are not readily available from an existing source, the additional time and cost involved in obtaining the data must be taken into account. In all cases, the decision maker should

FIGURE 1.4 CUSTOMER OPINION QUESTIONNAIRE USED BY CHOPS CITY GRILL RESTAURANT IN NAPLES, FLORIDA



Chop/s
CITY GRILL

Date: ____________ Server Name: ____________

*O*ur customers are our top priority. Please take a moment to fill out our survey card, so we can better serve your needs. You may return this card to the front desk or return by mail. Thank you!

SERVICE SURVEY	Excellent	Good	Average	Fair	Poor
Overall Experience	□	□	□	□	□
Greeting by Hostess	□	□	□	□	□
Manager (Table Visit)	□	□	□	□	□
Overall Service	□	□	□	□	□
Professionalism	□	□	□	□	□
Menu Knowledge	□	□	□	□	□
Friendliness	□	□	□	□	□
Wine Selection	□	□	□	□	□
Menu Selection	□	□	□	□	□
Food Quality	□	□	□	□	□
Food Presentation	□	□	□	□	□
Value for $ Spent	□	□	□	□	□

What comments could you give us to improve our restaurant?

____

____

Thank you, we appreciate your comments. —The staff of Chops City Grill.

consider the contribution of the statistical analysis to the decision-making process. The cost of data acquisition and the subsequent statistical analysis should not exceed the savings generated by using the information to make a better decision.

Data Acquisition Errors

Managers should always be aware of the possibility of data errors in statistical studies. Using erroneous data can be worse than not using any data at all. An error in data acquisition occurs whenever the data value obtained is not equal to the true or actual value that would be obtained with a correct procedure. Such errors can occur in a number of ways. For example, an interviewer might make a recording error, such as a transposition in writing the age of a 24-year-old person as 42, or the person answering an interview question might misinterpret the question and provide an incorrect response.

Experienced data analysts take great care in collecting and recording data to ensure that errors are not made. Special procedures can be used to check for internal consistency of the data. For instance, such procedures would indicate that the analyst should review the accuracy of data for a respondent shown to be 22 years of age but reporting 20 years of work experience. Data analysts also review data with unusually large and small values, called outliers, which are candidates for possible data errors. In Chapter 3 we present some of the methods statisticians use to identify outliers.

Errors often occur during data acquisition. Blindly using any data that happen to be available or using data that were acquired with little care can result in misleading information and bad decisions. Thus, taking steps to acquire accurate data can help ensure reliable and valuable decision-making information.

1.4 Descriptive Statistics

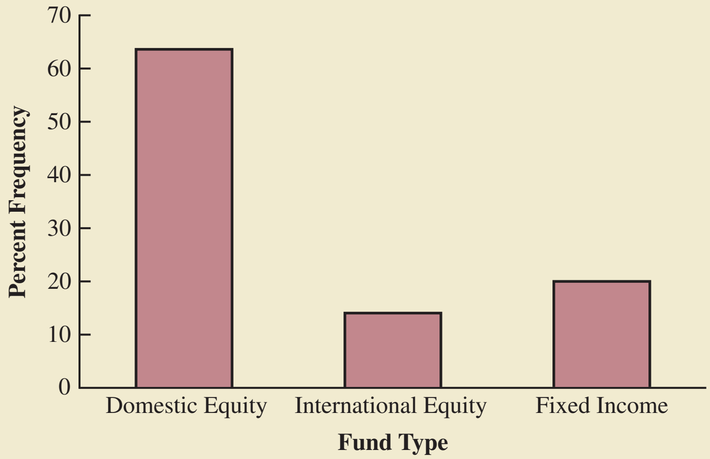
Most of the statistical information in newspapers, magazines, company reports, and other publications consists of data that are summarized and presented in a form that is easy for the reader to understand. Such summaries of data, which may be tabular, graphical, or numerical, are referred to as **descriptive statistics**.

Refer again to the data set in Table 1.1 showing data on 25 mutual funds. Methods of descriptive statistics can be used to provide summaries of the information in this data set. For example, a tabular summary of the data for the categorical variable Fund Type is shown in Table 1.4. A graphical summary of the same data, called a bar chart, is shown in Figure 1.5. These types of tabular and graphical summaries generally make the data easier to interpret. Referring to Table 1.4 and Figure 1.5, we can see easily that the majority of the mutual funds are of the Domestic Equity type. On a percentage basis, 64% are of the Domestic Equity type, 16% are of the International Equity type, and 20% are of the Fixed Income type.

TABLE 1.4 FREQUENCIES AND PERCENT FREQUENCIES FOR MUTUAL FUND TYPE

Mutual Fund Type	Frequency	Percent Frequency
Domestic Equity	16	64
International Equity	4	16
Fixed Income	5	20
Totals	**25**	**100**

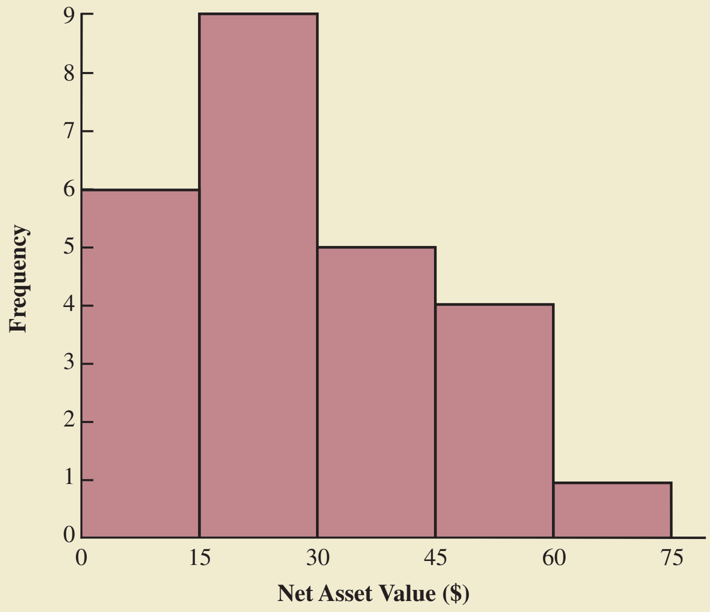
FIGURE 1.5 BAR CHART FOR MUTUAL FUND TYPE



A graphical summary of the data for the quantitative variable Net Asset Value, called a histogram, is provided in Figure 1.6. The histogram makes it easy to see that the net asset values range from $0 to $75, with the highest concentration between $15 and $30. Only one of the net asset values is greater than $60.

In addition to tabular and graphical displays, numerical descriptive statistics are used to summarize data. The most common numerical descriptive statistic is the average, or

FIGURE 1.6 HISTOGRAM OF NET ASSET VALUE FOR 25 MUTUAL FUNDS



mean. Using the data on 5-Year Average Return for the mutual funds in Table 1.1, we can compute the average by adding the returns for all 25 mutual funds and dividing the sum by 25. Doing so provides a 5-year average return of 16.50%. This average demonstrates a measure of the central tendency, or central location, of the data for that variable.

There is a great deal of interest in effective methods for developing and presenting descriptive statistics. Chapters 2 and 3 devote attention to the tabular, graphical, and numerical methods of descriptive statistics.

1.5 Statistical Inference

Many situations require information about a large group of elements (individuals, companies, voters, households, products, customers, and so on). But, because of time, cost, and other considerations, data can be collected from only a small portion of the group. The larger group of elements in a particular study is called the **population**, and the smaller group is called the **sample**. Formally, we use the following definitions.

POPULATION

A population is the set of all elements of interest in a particular study.

SAMPLE

A sample is a subset of the population.

The U.S. government conducts a census every 10 years. Market research firms conduct sample surveys every day.

The process of conducting a survey to collect data for the entire population is called a **census**. The process of conducting a survey to collect data for a sample is called a **sample survey**. As one of its major contributions, statistics uses data from a sample to make estimates and test hypotheses about the characteristics of a population through a process referred to as **statistical inference**.

As an example of statistical inference, let us consider the study conducted by Norris Electronics. Norris manufactures a high-intensity lightbulb used in a variety of electrical products. In an attempt to increase the useful life of the lightbulb, the product design group developed a new lightbulb filament. In this case, the population is defined as all lightbulbs that could be produced with the new filament. To evaluate the advantages of the new filament, 200 bulbs with the new filament were manufactured and tested. Data collected from this sample showed the number of hours each lightbulb operated before filament burnout. See Table 1.5.

Suppose Norris wants to use the sample data to make an inference about the average hours of useful life for the population of all lightbulbs that could be produced with the new filament. Adding the 200 values in Table 1.5 and dividing the total by 200 provides the sample average lifetime for the lightbulbs: 76 hours. We can use this sample result to estimate that the average lifetime for the lightbulbs in the population is 76 hours. Figure 1.7 provides a graphical summary of the statistical inference process for Norris Electronics.

Whenever statisticians use a sample to estimate a population characteristic of interest, they usually provide a statement of the quality, or precision, associated with the estimate.

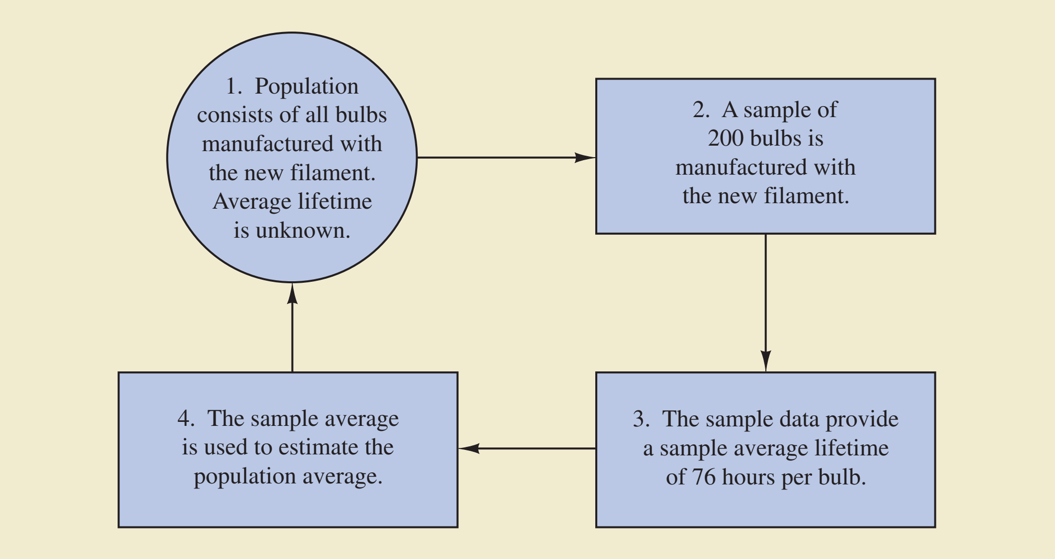
TABLE 1.5 HOURS UNTIL BURNOUT FOR A SAMPLE OF 200 LIGHTBULBS FOR THE NORRIS ELECTRONICS EXAMPLE

WEB file

Norris

107	73	68	97	76	79	94	59	98	57
54	65	71	70	84	88	62	61	79	98
66	62	79	86	68	74	61	82	65	98
62	116	65	88	64	79	78	79	77	86
74	85	73	80	68	78	89	72	58	69
92	78	88	77	103	88	63	68	88	81
75	90	62	89	71	71	74	70	74	70
65	81	75	62	94	71	85	84	83	63
81	62	79	83	93	61	65	62	92	65
83	70	70	81	77	72	84	67	59	58
78	66	66	94	77	63	66	75	68	76
90	78	71	101	78	43	59	67	61	71
96	75	64	76	72	77	74	65	82	86
66	86	96	89	81	71	85	99	59	92
68	72	77	60	87	84	75	77	51	45
85	67	87	80	84	93	69	76	89	75
83	68	72	67	92	89	82	96	77	102
74	91	76	83	66	68	61	73	72	76
73	77	79	94	63	59	62	71	81	65
73	63	63	89	82	64	85	92	64	73

FIGURE 1.7 THE PROCESS OF STATISTICAL INFERENCE FOR THE NORRIS ELECTRONICS EXAMPLE



1. Population consists of all bulbs manufactured with the new filament. Average lifetime is unknown.

2. A sample of 200 bulbs is manufactured with the new filament.

3. The sample data provide a sample average lifetime of 76 hours per bulb.

4. The sample average is used to estimate the population average.

For the Norris example, the statistician might state that the point estimate of the average lifetime for the population of new lightbulbs is 76 hours with a margin of error of ±4 hours. Thus, an interval estimate of the average lifetime for all lightbulbs produced with the new filament is 72 hours to 80 hours. The statistician can also state how confident he or she is that the interval from 72 hours to 80 hours contains the population average.

1.6 Computers and Statistical Analysis

Statisticians frequently use computer software to perform the statistical computations required with large amounts of data. For example, computing the average lifetime for the 200 lightbulbs in the Norris Electronics example (see Table 1.5) would be quite tedious without a computer. To facilitate computer usage, many of the data sets in this book are available on the website that accompanies the text. The data files may be downloaded in either Minitab or Excel formats. In addition, the Excel add-in StatTools can be downloaded from the website. End-of-chapter appendixes cover the step-by-step procedures for using Minitab, Excel, and the Excel add-in StatTools to implement the statistical techniques presented in the chapter.

Minitab and Excel data sets and the Excel add-in StatTools are available on the website for this text.

1.7 Data Mining

With the aid of magnetic card readers, bar code scanners, and point-of-sale terminals, most organizations obtain large amounts of data on a daily basis. And, even for a small local restaurant that uses touch screen monitors to enter orders and handle billing, the amount of data collected can be significant. For large retail companies, the sheer volume of data collected is hard to conceptualize, and figuring out how to effectively use these data to improve profitability is a challenge. For example, mass retailers such as Wal-Mart capture data on 20 to 30 million transactions every day, telecommunication companies such as France Telecom and AT&T generate over 300 million call records per day, and Visa processes 6800 payment transactions per second or approximately 600 million transactions per day. Storing and managing the transaction data is a significant undertaking.

The term *data warehousing* is used to refer to the process of capturing, storing, and maintaining the data. Computing power and data collection tools have reached the point where it is now feasible to store and retrieve extremely large quantities of data in seconds. Analysis of the data in the warehouse may result in decisions that will lead to new strategies and higher profits for the organization.

The subject of **data mining** deals with methods for developing useful decision-making information from large data bases. Using a combination of procedures from statistics, mathematics, and computer science, analysts "mine the data" in the warehouse to convert it into useful information, hence the name *data mining*. Dr. Kurt Thearling, a leading practitioner in the field, defines data mining as "the automated extraction of predictive information from (large) databases." The two key words in Dr. Thearling's definition are "automated" and "predictive." Data mining systems that are the most effective use automated procedures to extract information from the data using only the most general or even vague queries by the user. And data mining software automates the process of uncovering hidden predictive information that in the past required hands-on analysis.

The major applications of data mining have been made by companies with a strong consumer focus, such as retail businesses, financial organizations, and communication companies. Data mining has been successfully used to help retailers such as Amazon and Barnes & Noble determine one or more related products that customers who have already purchased a specific product are also likely to purchase. Then, when a customer logs on to the company's website and purchases a product, the website uses pop-ups to alert the customer about additional products that the customer is likely to purchase. In another application, data mining may be used to identify customers who are likely to spend more than $20 on a particular shopping trip. These customers may then be identified as the ones to receive special e-mail or regular mail discount offers to encourage them to make their next shopping trip before the discount termination date.

Data mining is a technology that relies heavily on statistical methodology such as multiple regression, logistic regression, and correlation. But it takes a creative integration of all

Statistical methods play an important role in data mining, both in terms of discovering relationships in the data and predicting future outcomes. However, a thorough coverage of data mining and the use of statistics in data mining are outside the scope of this text.

these methods and computer science technologies involving artificial intelligence and machine learning to make data mining effective. A significant investment in time and money is required to implement commercial data mining software packages developed by firms such as Oracle, Teradata, and SAS. The statistical concepts introduced in this text will be helpful in understanding the statistical methodology used by data mining software packages and enable you to better understand the statistical information that is developed.

Because statistical models play an important role in developing predictive models in data mining, many of the concerns that statisticians deal with in developing statistical models are also applicable. For instance, a concern in any statistical study involves the issue of model reliability. Finding a statistical model that works well for a particular sample of data does not necessarily mean that it can be reliably applied to other data. One of the common statistical approaches to evaluating model reliability is to divide the sample data set into two parts: a training data set and a test data set. If the model developed using the training data is able to accurately predict values in the test data, we say that the model is reliable. One advantage that data mining has over classical statistics is that the enormous amount of data available allows the data mining software to partition the data set so that a model developed for the training data set may be tested for reliability on other data. In this sense, the partitioning of the data set allows data mining to develop models and relationships and then quickly observe if they are repeatable and valid with new and different data. On the other hand, a warning for data mining applications is that with so much data available, there is a danger of overfitting the model to the point that misleading associations and cause/effect conclusions appear to exist. Careful interpretation of data mining results and additional testing will help avoid this pitfall.

1.8 Ethical Guidelines for Statistical Practice

Ethical behavior is something we should strive for in all that we do. Ethical issues arise in statistics because of the important role statistics plays in the collection, analysis, presentation, and interpretation of data. In a statistical study, unethical behavior can take a variety of forms including improper sampling, inappropriate analysis of the data, development of misleading graphs, use of inappropriate summary statistics, and/or a biased interpretation of the statistical results.

As you begin to do your own statistical work, we encourage you to be fair, thorough, objective, and neutral as you collect data, conduct analyses, make oral presentations, and present written reports containing information developed. As a consumer of statistics, you should also be aware of the possibility of unethical statistical behavior by others. When you see statistics in newspapers, on television, on the Internet, and so on, it is a good idea to view the information with some skepticism, always being aware of the source as well as the purpose and objectivity of the statistics provided.

The American Statistical Association, the nation’s leading professional organization for statistics and statisticians, developed the report “Ethical Guidelines for Statistical Practice”[1] to help statistical practitioners make and communicate ethical decisions and assist students in learning how to perform statistical work responsibly. The report contains 67 guidelines organized into eight topic areas: Professionalism; Responsibilities to Funders, Clients, and Employers; Responsibilities in Publications and Testimony; Responsibilities to Research Subjects; Responsibilities to Research Team Colleagues; Responsibilities to Other Statisticians or Statistical Practitioners; Responsibilities Regarding Allegations of Misconduct; and Responsibilities of Employers Including Organizations, Individuals, Attorneys, or Other Clients Employing Statistical Practitioners.

[1] American Statistical Association “Ethical Guidelines for Statistical Practice,” 1999.

One of the ethical guidelines in the professionalism area addresses the issue of running multiple tests until a desired result is obtained. Let us consider an example. In Section 1.5 we discussed a statistical study conducted by Norris Electronics involving a sample of 200 high-intensity lightbulbs manufactured with a new filament. The average lifetime for the sample, 76 hours, provided an estimate of the average lifetime for all lightbulbs produced with the new filament. However, consider this. Because Norris selected a sample of bulbs, it is reasonable to assume that another sample would have provided a different average lifetime.

Suppose Norris's management had hoped the sample results would enable them to claim that the average lifetime for the new lightbulbs was 80 hours or more. Suppose further that Norris's management decides to continue the study by manufacturing and testing repeated samples of 200 lightbulbs with the new filament until a sample mean of 80 hours or more is obtained. If the study is repeated enough times, a sample may eventually be obtained—by chance alone—that would provide the desired result and enable Norris to make such a claim. In this case, consumers would be misled into thinking the new product is better than it actually is. Clearly, this type of behavior is unethical and represents a gross misuse of statistics in practice.

Several ethical guidelines in the responsibilities and publications and testimony area deal with issues involving the handling of data. For instance, a statistician must account for all data considered in a study and explain the sample(s) actually used. In the Norris Electronics study the average lifetime for the 200 bulbs in the original sample is 76 hours; this is considerably less than the 80 hours or more that management hoped to obtain. Suppose now that after reviewing the results showing a 76 hour average lifetime, Norris discards all the observations with 70 or fewer hours until burnout, allegedly because these bulbs contain imperfections caused by startup problems in the manufacturing process. After discarding these lightbulbs, the average lifetime for the remaining lightbulbs in the sample turns out to be 82 hours. Would you be suspicious of Norris's claim that the lifetime for their lightbulbs is 82 hours?

If the Norris lightbulbs showing 70 or fewer hours until burnout were discarded to simply provide an average lifetime of 82 hours, there is no question that discarding the lightbulbs with 70 or fewer hours until burnout is unethical. But, even if the discarded lightbulbs contain imperfections due to startup problems in the manufacturing process—and, as a result, should not have been included in the analysis—the statistician who conducted the study must account for all the data that were considered and explain how the sample actually used was obtained. To do otherwise is potentially misleading and would constitute unethical behavior on the part of both the company and the statistician.

A guideline in the shared values section of the American Statistical Association report states that statistical practitioners should avoid any tendency to slant statistical work toward predetermined outcomes. This type of unethical practice is often observed when unrepresentative samples are used to make claims. For instance, in many areas of the country smoking is not permitted in restaurants. Suppose, however, a lobbyist for the tobacco industry interviews people in restaurants where smoking is permitted in order to estimate the percentage of people who are in favor of allowing smoking in restaurants. The sample results show that 90% of the people interviewed are in favor of allowing smoking in restaurants. Based upon these sample results, the lobbyist claims that 90% of all people who eat in restaurants are in favor of permitting smoking in restaurants. In this case we would argue that only sampling persons eating in restaurants that allow smoking has biased the results. If only the final results of such a study are reported, readers unfamiliar with the details of the study (i.e., that the sample was collected only in restaurants allowing smoking) can be misled.

The scope of the American Statistical Association's report is broad and includes ethical guidelines that are appropriate not only for a statistician, but also for consumers of statistical information. We encourage you to read the report to obtain a better perspective of ethical issues as you continue your study of statistics and to gain the background for determining how to ensure that ethical standards are met when you start to use statistics in practice.

Summary

Statistics is the art and science of collecting, analyzing, presenting, and interpreting data. Nearly every college student majoring in business or economics is required to take a course in statistics. We began the chapter by describing typical statistical applications for business and economics.

Data consist of the facts and figures that are collected and analyzed. Four scales of measurement used to obtain data on a particular variable include nominal, ordinal, interval, and ratio. The scale of measurement for a variable is nominal when the data are labels or names used to identify an attribute of an element. The scale is ordinal if the data demonstrate the properties of nominal data and the order or rank of the data is meaningful. The scale is interval if the data demonstrate the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure. Finally, the scale of measurement is ratio if the data show all the properties of interval data and the ratio of two values is meaningful.

For purposes of statistical analysis, data can be classified as categorical or quantitative. Categorical data use labels or names to identify an attribute of each element. Categorical data use either the nominal or ordinal scale of measurement and may be nonnumeric or numeric. Quantitative data are numeric values that indicate how much or how many. Quantitative data use either the interval or ratio scale of measurement. Ordinary arithmetic operations are meaningful only if the data are quantitative. Therefore, statistical computations used for quantitative data are not always appropriate for categorical data.

In Sections 1.4 and 1.5 we introduced the topics of descriptive statistics and statistical inference. Descriptive statistics are the tabular, graphical, and numerical methods used to summarize data. The process of statistical inference uses data obtained from a sample to make estimates or test hypotheses about the characteristics of a population. The last three sections of the chapter provide information on the role of computers in statistical analysis, an introduction to the relative new field of data mining, and a summary of ethical guidelines for statistical practice.

Glossary

Statistics The art and science of collecting, analyzing, presenting, and interpreting data.

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

Data set All the data collected in a particular study.

Elements The entities on which data are collected.

Variable A characteristic of interest for the elements.

Observation The set of measurements obtained for a particular element.

Nominal scale The scale of measurement for a variable when the data are labels or names used to identify an attribute of an element. Nominal data may be nonnumeric or numeric.

Ordinal scale The scale of measurement for a variable if the data exhibit the properties of nominal data and the order or rank of the data is meaningful. Ordinal data may be nonnumeric or numeric.

Interval scale The scale of measurement for a variable if the data demonstrate the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure. Interval data are always numeric.

Ratio scale The scale of measurement for a variable if the data demonstrate all the properties of interval data and the ratio of two values is meaningful. Ratio data are always numeric.

Categorical data Labels or names used to identify an attribute of each element. Categorical data use either the nominal or ordinal scale of measurement and may be nonnumeric or numeric.
Quantitative data Numeric values that indicate how much or how many of something. Quantitative data are obtained using either the interval or ratio scale of measurement.
Categorical variable A variable with categorical data.
Quantitative variable A variable with quantitative data.
Cross-sectional data Data collected at the same or approximately the same point in time.
Time series data Data collected over several time periods.
Descriptive statistics Tabular, graphical, and numerical summaries of data.
Population The set of all elements of interest in a particular study.
Sample A subset of the population.
Census A survey to collect data on the entire population.
Sample survey A survey to collect data on a sample.
Statistical inference The process of using data obtained from a sample to make estimates or test hypotheses about the characteristics of a population.
Data mining The process of using procedures from statistics and computer science to extract useful information from extremely large databases.

Supplementary Exercises

1. Discuss the differences between statistics as numerical facts and statistics as a discipline or field of study.

SELF test

2. The U.S. Department of Energy provides fuel economy information for a variety of motor vehicles. A sample of 10 automobiles is shown in Table 1.6 (Fuel Economy website, February 22, 2008). Data show the size of the automobile (compact, midsize, or large), the number of cylinders in the engine, the city driving miles per gallon, the highway driving miles per gallon, and the recommended fuel (diesel, premium, or regular).
   a. How many elements are in this data set?
   b. How many variables are in this data set?
   c. Which variables are categorical and which variables are quantitative?
   d. What type of measurement scale is used for each of the variables?

SELF test

3. Refer to Table 1.6.
   a. What is the average miles per gallon for city driving?
   b. On average, how much higher is the miles per gallon for highway driving as compared to city driving?

TABLE 1.6 FUEL ECONOMY INFORMATION FOR 10 AUTOMOBILES

Car	Size	Cylinders	City MPG	Highway MPG	Fuel
Audi A8	Large	12	13	19	Premium
BMW 328Xi	Compact	6	17	25	Premium
Cadillac CTS	Midsize	6	16	25	Regular
Chrysler 300	Large	8	13	18	Premium
Ford Focus	Compact	4	24	33	Regular
Hyundai Elantra	Midsize	4	25	33	Regular
Jeep Grand Cherokee	Midsize	6	17	26	Diesel
Pontiac G6	Compact	6	15	22	Regular
Toyota Camry	Midsize	4	21	31	Regular
Volkswagen Jetta	Compact	5	21	29	Regular

TABLE 1.7 DATA FOR SEVEN COLLEGES AND UNIVERSITIES

				%	
		Campus	**Endowment**	**Applicants**	**NCAA**
School	**State**	**Setting**	**($ billions)**	**Admitted**	**Division**
Amherst College	Massachusetts	Town: Fringe	1.7	18	III
Duke	North Carolina	City: Midsize	5.9	21	I-A
Harvard University	Massachusetts	City: Midsize	34.6	9	I-AA
Swarthmore College	Pennsylvania	Suburb: Large	1.4	18	III
University of Pennsylvania	Pennsylvania	City: Large	6.6	18	I-AA
Williams College	Massachusetts	Town: Fringe	1.9	18	III
Yale University	Connecticut	City: Midsize	22.5	9	I-AA

c. What percentage of the cars have four-cylinder engines?
d. What percentage of the cars use regular fuel?

4. Table 1.7 shows data for seven colleges and universities. The endowment (in billions of dollars) and the percentage of applicants admitted are shown (*USA Today,* February 3, 2008). The state each school is located in, the campus setting, and the NCAA Division for varsity teams were obtained from the National Center of Education Statistics website, February 22, 2008.
   a. How many elements are in the data set?
   b. How many variables are in the data set?
   c. Which of the variables are categorical and which are quantitative?

5. Consider the data set in Table 1.7
   a. Compute the average endowment for the sample.
   b. Compute the average percentage of applicants admitted.
   c. What percentage of the schools have NCAA Division III varsity teams?
   d. What percentage of the schools have a City: Midsize campus setting?

6. *Foreign Affairs* magazine conducted a survey to develop a profile of its subscribers (Foreign Affairs website, February 23, 2008). The following questions were asked.
   a. How many nights have you stayed in a hotel in the past 12 months?
   b. Where do you purchase books? Three options were listed: Bookstore, Internet, and Book Club.
   c. Do you own or lease a luxury vehicle? (Yes or No)
   d. What is your age?
   e. For foreign trips taken in the past three years, what was your destination? Seven international destinations were listed.

   Comment on whether each question provides categorical or quantitative data.

7. The Ritz-Carlton Hotel used a customer opinion questionnaire to obtain performance data about its dining and entertainment services (The Ritz-Carlton Hotel, Naples, Florida, February 2006). Customers were asked to rate six factors: Welcome, Service, Food, Menu Appeal, Atmosphere, and Overall Experience. Data were recorded for each factor with 1 for Fair, 2 for Average, 3 for Good, and 4 for Excellent.
   a. The customer responses provided data for six variables. Are the variables categorical or quantitative?
   b. What measurement scale is used?

8. The *FinancialTimes*/Harris Poll is a monthly online poll of adults from six countries in Europe and the United States. A January poll included 1015 adults in the United States. One of the questions asked was, "How would you rate the Federal Bank in handling the

credit problems in the financial markets?" Possible responses were Excellent, Good, Fair, Bad, and Terrible (Harris Interactive website, January 2008).

a. What was the sample size for this survey?
b. Are the data categorical or quantitative?
c. Would it make more sense to use averages or percentages as a summary of the data for this question?
d. Of the respondents in the United States, 10% said the Federal Bank is doing a good job. How many individuals provided this response?

9. The Commerce Department reported receiving the following applications for the Malcolm Baldrige National Quality Award: 23 from large manufacturing firms, 18 from large service firms, and 30 from small businesses.

a. Is type of business a categorical or quantitative variable?
b. What percentage of the applications came from small businesses?

10. *The Wall Street Journal (WSJ)* subscriber survey (October 13, 2003) asked 46 questions about subscriber characteristics and interests. State whether each of the following questions provided categorical or quantitative data and indicate the measurement scale appropriate for each.

a. What is your age?
b. Are you male or female?
c. When did you first start reading the *WSJ*? High school, college, early career, mid-career, late career, or retirement?
d. How long have you been in your present job or position?
e. What type of vehicle are you considering for your next purchase? Nine response categories include sedan, sports car, SUV, minivan, and so on.

11. State whether each of the following variables is categorical or quantitative and indicate its measurement scale.

a. Annual sales
b. Soft drink size (small, medium, large)
c. Employee classification (GS1 through GS18)
d. Earnings per share
e. Method of payment (cash, check, credit card)

12. The Hawaii Visitors Bureau collects data on visitors to Hawaii. The following questions were among 16 asked in a questionnaire handed out to passengers during incoming airline flights in June 2003.

- This trip to Hawaii is my: 1st, 2nd, 3rd, 4th, etc.
- The primary reason for this trip is: (10 categories including vacation, convention, honeymoon)
- Where I plan to stay: (11 categories including hotel, apartment, relatives, camping)
- Total days in Hawaii

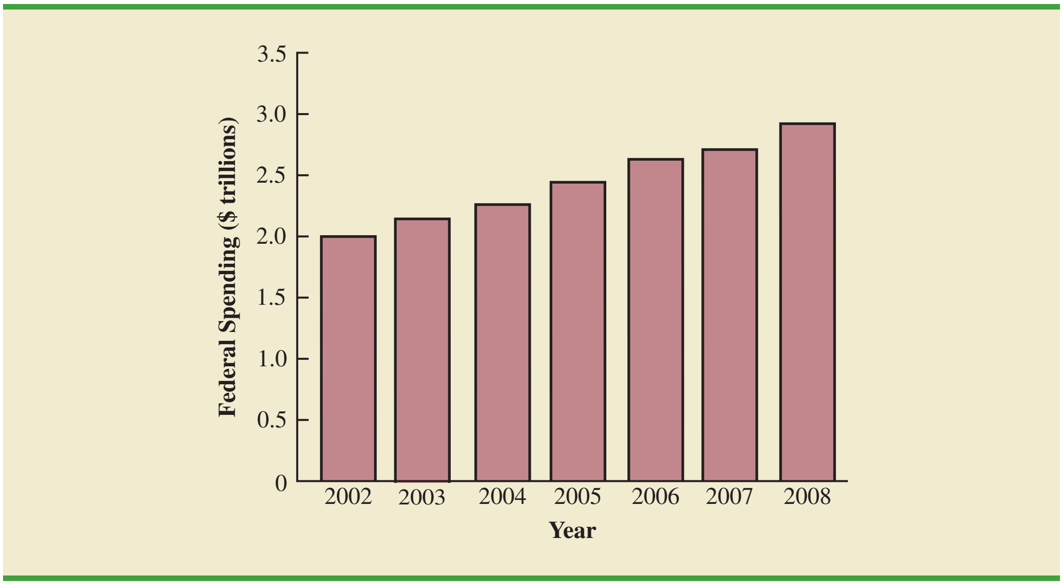
a. What is the population being studied?
b. Is the use of a questionnaire a good way to reach the population of passengers on incoming airline flights?
c. Comment on each of the four questions in terms of whether it will provide categorical or quantitative data.

SELF test

13. Figure 1.8 provides a bar chart showing the amount of federal spending for the years 2002 to 2008 (*USA Today,* February 5, 2008).

a. What is the variable of interest?
b. Are the data categorical or quantitative?
c. Are the data time series or cross-sectional?
d. Comment on the trend in federal spending over time.

FIGURE 1.8 FEDERAL SPENDING



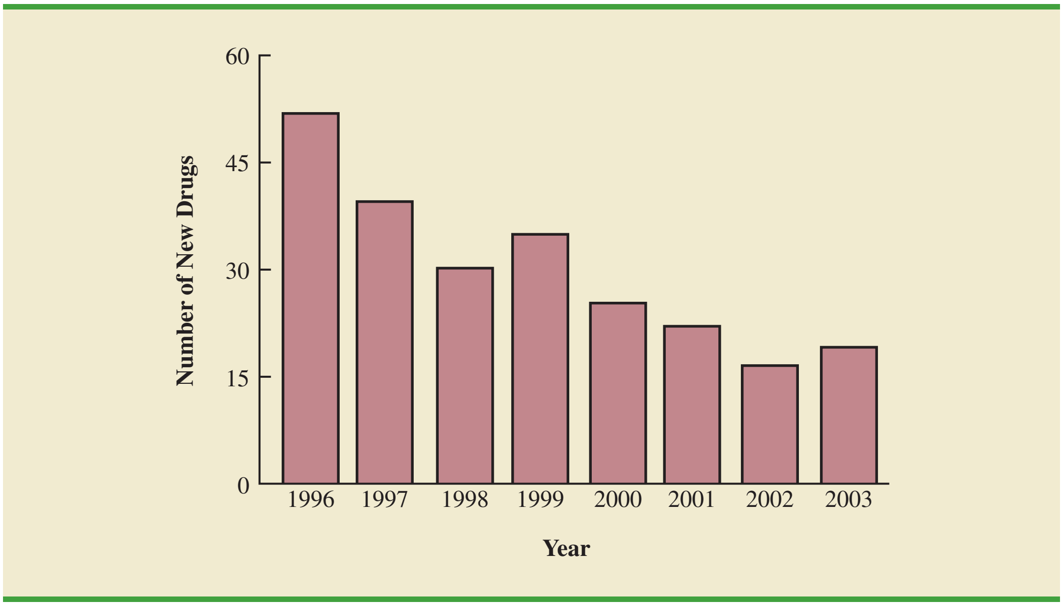
14. CSM Worldwide forecasts global production for all automobile manufacturers. The following CSM data show the forecast of global auto production for General Motors, Ford, DaimlerChrysler, and Toyota for the years 2004 to 2007 (*USA Today,* December 21, 2005). Data are in millions of vehicles.

Manufacturer	**2004**	**2005**	**2006**	**2007**
General Motors	8.9	9.0	8.9	8.8
Ford	7.8	7.7	7.8	7.9
DaimlerChrysler	4.1	4.2	4.3	4.6
Toyota	7.8	8.3	9.1	9.6

a. Construct a time series graph for the years 2004 to 2007 showing the number of vehicles manufactured by each automotive company. Show the time series for all four manufacturers on the same graph.
b. General Motors has been the undisputed production leader of automobiles since 1931. What does the time series graph show about who is the world's biggest car company? Discuss.
c. Construct a bar graph showing vehicles produced by automobile manufacturer using the 2007 data. Is this graph based on cross-sectional or time series data?

15. The Food and Drug Administration (FDA) reported the number of new drugs approved over an eight-year period (*The Wall Street Journal,* January 12, 2004). Figure 1.9 provides a bar chart summarizing the number of new drugs approved each year.
a. Are the data categorical or quantitative?
b. Are the data time series or cross-sectional?
c. How many new drugs were approved in 2003?
d. In what year were the fewest new drugs approved? How many?
e. Comment on the trend in the number of new drugs approved by the FDA over the eight-year period.

FIGURE 1.9 NUMBER OF NEW DRUGS APPROVED BY THE FOOD AND DRUG ADMINISTRATION



16. The Energy Information Administration of the U.S. Department of Energy provided time series data for the U.S. average price per gallon of conventional regular gasoline between July 2006 and June 2009 (Energy Information Administration website, June 2009). Use the Internet to obtain the average price per gallon of conventional regular gasoline since June 2009.
    a. Extend the graph of the time series shown in Figure 1.1.
    b. What interpretations can you make about the average price per gallon of conventional regular gasoline since June 2009?
    c. Does the time series continue to show a summer increase in the average price per gallon? Explain.

17. A manager of a large corporation recommends a $10,000 raise be given to keep a valued subordinate from moving to another company. What internal and external sources of data might be used to decide whether such a salary increase is appropriate?

18. A survey of 430 business travelers found 155 business travelers used a travel agent to make the travel arrangements (*USA Today,* November 20, 2003).
    a. Develop a descriptive statistic that can be used to estimate the percentage of all business travelers who use a travel agent to make travel arrangements.
    b. The survey reported that the most frequent way business travelers make travel arrangements is by using an online travel site. If 44% of business travelers surveyed made their arrangements this way, how many of the 430 business travelers used an online travel site?
    c. Are the data on how travel arrangements are made categorical or quantitative?

19. A *BusinessWeek* North American subscriber study collected data from a sample of 2861 subscribers. Fifty-nine percent of the respondents indicated an annual income of $75,000 or more, and 50% reported having an American Express credit card.
    a. What is the population of interest in this study?
    b. Is annual income a categorical or quantitative variable?
    c. Is ownership of an American Express card a categorical or quantitative variable?
    d. Does this study involve cross-sectional or time series data?
    e. Describe any statistical inferences *BusinessWeek* might make on the basis of the survey.

20. A survey of 131 investment managers in *Barron's* Big Money poll revealed the following:
    - 43% of managers classified themselves as bullish or very bullish on the stock market.
    - The average expected return over the next 12 months for equities was 11.2%.
    - 21% selected health care as the sector most likely to lead the market in the next 12 months.
    - When asked to estimate how long it would take for technology and telecom stocks to resume sustainable growth, the managers' average response was 2.5 years.

    a. Cite two descriptive statistics.
    b. Make an inference about the population of all investment managers concerning the average return expected on equities over the next 12 months.
    c. Make an inference about the length of time it will take for technology and telecom stocks to resume sustainable growth.

21. A seven-year medical research study reported that women whose mothers took the drug DES during pregnancy were twice as likely to develop tissue abnormalities that might lead to cancer as were women whose mothers did not take the drug.
    a. This study involved the comparison of two populations. What were the populations?
    b. Do you suppose the data were obtained in a survey or an experiment?
    c. For the population of women whose mothers took the drug DES during pregnancy, a sample of 3980 women showed 63 developed tissue abnormalities that might lead to cancer. Provide a descriptive statistic that could be used to estimate the number of women out of 1000 in this population who have tissue abnormalities.
    d. For the population of women whose mothers did not take the drug DES during pregnancy, what is the estimate of the number of women out of 1000 who would be expected to have tissue abnormalities?
    e. Medical studies often use a relatively large sample (in this case, 3980). Why?

22. The Nielsen Company surveyed consumers in 47 markets from Europe, Asia-Pacific, the Americas, and the Middle East to determine which factors are most important in determining where they buy groceries. Using a scale of 1 (low) to 5 (high), the highest rated factor was *good value for money*, with an average point score of 4.32. The second highest rated factor was *better selection of high-quality brands and products,* with an average point score of 3.78, and the lowest rated factor was *uses recyclable bags and packaging*, with an average point score of 2.71 (Nielsen website, February 24, 2008). Suppose that you have been hired by a grocery store chain to conduct a similar study to determine what factors customers at the chain's stores in Charlotte, North Carolina, think are most important in determining where they buy groceries.
    a. What is the population for the survey that you will be conducting?
    b. How would you collect the data for this study?

23. Nielsen Media Research conducts weekly surveys of television viewing throughout the United States, publishing both rating and market share data. The Nielsen rating is the percentage of households with televisions watching a program, while the Nielsen share is the percentage of households watching a program among those households with televisions in use. For example, Nielsen Media Research results for the 2003 Baseball World Series between the New York Yankees and the Florida Marlins showed a rating of 12.8% and a share of 22% (Associated Press, October 27, 2003). Thus, 12.8% of households with televisions were watching the World Series and 22% of households with televisions in use were watching the World Series. Based on the rating and share data for major television programs, Nielsen publishes a weekly ranking of television programs as well as a weekly ranking of the four major networks: ABC, CBS, NBC, and Fox.
    a. What is Nielsen Media Research attempting to measure?
    b. What is the population?
    c. Why would a sample be used in this situation?
    d. What kinds of decisions or actions are based on the Nielsen rankings?

TABLE 1.8 DATA SET FOR 25 SHADOW STOCKS

WEB file

Shadow02

Company	Exchange	Ticker Symbol	Market Cap ($ millions)	Price/ Earnings Ratio	Gross Profit Margin (%)
DeWolfe Companies	AMEX	DWL	36.4	8.4	36.7
North Coast Energy	OTC	NCEB	52.5	6.2	59.3
Hansen Natural Corp.	OTC	HANS	41.1	14.6	44.8
MarineMax, Inc.	NYSE	HZO	111.5	7.2	23.8
Nanometrics Incorporated	OTC	NANO	228.6	38.0	53.3
TeamStaff, Inc.	OTC	TSTF	92.1	33.5	4.1
Environmental Tectonics	AMEX	ETC	51.1	35.8	35.9
Measurement Specialties	AMEX	MSS	101.8	26.8	37.6
SEMCO Energy, Inc.	NYSE	SEN	193.4	18.7	23.6
Party City Corporation	OTC	PCTY	97.2	15.9	36.4
Embrex, Inc.	OTC	EMBX	136.5	18.9	59.5
Tech/Ops Sevcon, Inc.	AMEX	TO	23.2	20.7	35.7
ARCADIS NV	OTC	ARCAF	173.4	8.8	9.6
Qiao Xing Universal Tele.	OTC	XING	64.3	22.1	30.8
Energy West Incorporated	OTC	EWST	29.1	9.7	16.3
Barnwell Industries, Inc.	AMEX	BRN	27.3	7.4	73.4
Innodata Corporation	OTC	INOD	66.1	11.0	29.6
Medical Action Industries	OTC	MDCI	137.1	26.9	30.6
Instrumentarium Corp.	OTC	INMRY	240.9	3.6	52.1
Petroleum Development	OTC	PETD	95.9	6.1	19.4
Drexler Technology Corp.	OTC	DRXR	233.6	45.6	53.6
Gerber Childrenswear Inc.	NYSE	GCW	126.9	7.9	25.8
Gaiam, Inc.	OTC	GAIA	295.5	68.2	60.7
Artesian Resources Corp.	OTC	ARTNA	62.8	20.5	45.5
York Water Company	OTC	YORW	92.2	22.9	74.2

24. A sample of midterm grades for five students showed the following results: 72, 65, 82, 90, 76. Which of the following statements are correct, and which should be challenged as being too generalized?
    a. The average midterm grade for the sample of five students is 77.
    b. The average midterm grade for all students who took the exam is 77.
    c. An estimate of the average midterm grade for all students who took the exam is 77.
    d. More than half of the students who take this exam will score between 70 and 85.
    e. If five other students are included in the sample, their grades will be between 65 and 90.

25. Table 1.8 shows a data set containing information for 25 of the shadow stocks tracked by the American Association of Individual Investors. Shadow stocks are common stocks of smaller companies that are not closely followed by Wall Street analysts. The data set is also on the website that accompanies the text in the file named Shadow02.
    a. How many variables are in the data set?
    b. Which of the variables are categorical and which are quantitative?
    c. For the Exchange variable, show the frequency and the percent frequency for AMEX, NYSE, and OTC. Construct a bar graph similar to Figure 1.5 for the Exchange variable.
    d. Show the frequency distribution for the Gross Profit Margin using the five intervals: 0–14.9, 15–29.9, 30–44.9, 45–59.9, and 60–74.9. Construct a histogram similar to Figure 1.6.
    e. What is the average price/earnings ratio?

Appendix An Introduction to StatTools

StatTools is a professional add-in that expands the statistical capabilities available with Microsoft Excel. StatTools software can be downloaded from the website that accompanies this text.

Excel does not contain statistical functions or data analysis tools to perform all the statistical procedures discussed in the text. StatTools is a Microsoft Excel statistics add-in that extends the range of statistical and graphical options for Excel users. Most chapters include a chapter appendix that shows the steps required to accomplish a statistical procedure using StatTools. For those students who want to make more extensive use of the software, StatTools offers an excellent Help facility. The StatTools Help system includes detailed explanations of the statistical and data analysis options available, as well as descriptions and definitions of the types of output provided.

Getting Started with StatTools

StatTools software may be downloaded and installed on your computer by accessing the website that accompanies this text. After downloading and installing the software, perform the following steps to use StatTools as an Excel add-in.

Step 1. Click the **Start** button on the taskbar and then point to **All Programs**
Step 2. Point to the folder entitled **Palisade Decision Tools**
Step 3. Click **StatTools for Excel**

These steps will open Excel and add the StatTools tab next to the Add-Ins tab on the Excel Ribbon. Alternately, if you are already working in Excel, these steps will make StatTools available.

Using StatTools

Before conducting any statistical analysis, we must create a StatTools data set using the StatTools Data Set Manager. Let us use the Excel worksheet for the mutual funds data set in Table 1.1 to show how this is done. The following steps show how to create a StatTools data set for the mutual funds data.

Step 1. Open the Excel file named Morningstar
Step 2. Select any cell in the data set (for example, cell A1)
Step 3. Click the **StatTools** tab on the Ribbon
Step 4. In the **Data** group, click **Data Set Manager**
Step 5. When StatTools asks if you want to add the range $A$1:$F$26 as a new StatTools data set, click **Yes**
Step 6. When the StatTools—Data Set Manager dialog box appears, click **OK**

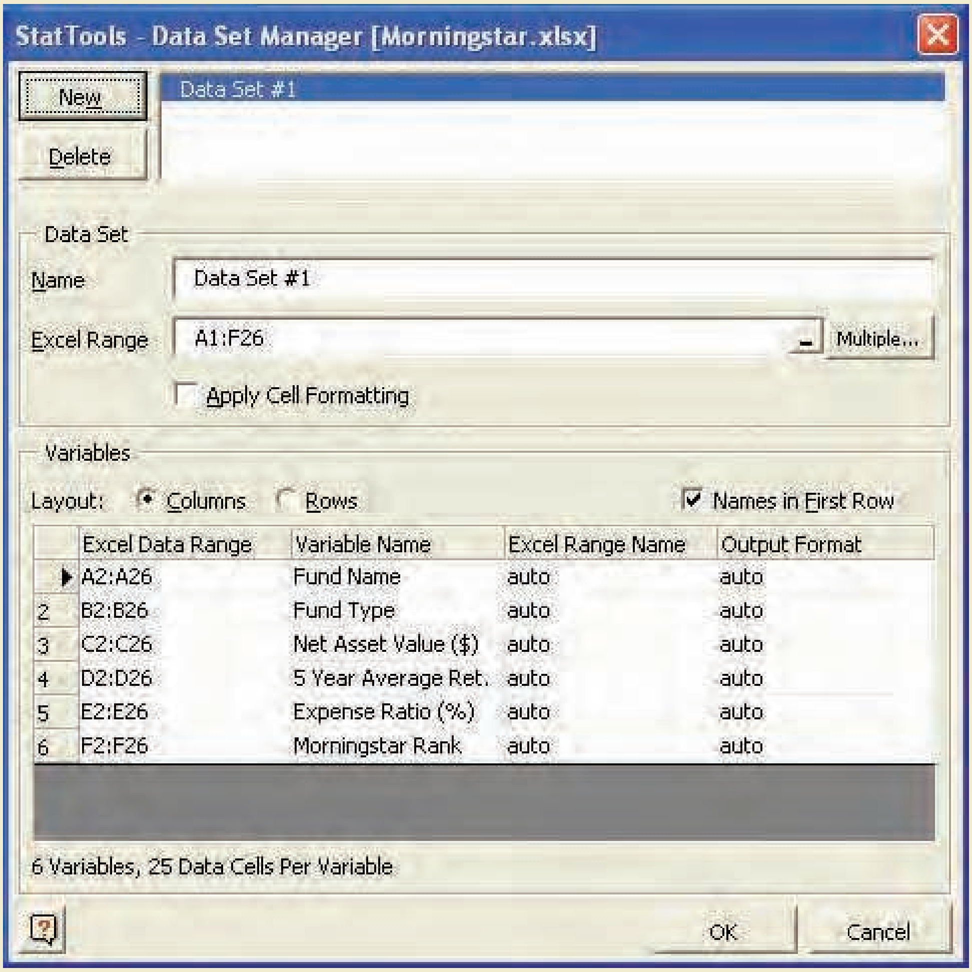
Figure 1.10 shows the StatTools—Data Set Manager dialog box that appears in step 6. By default, the name of the new StatTools data set is Data Set #1. You can replace the name Data Set #1 in step 6 with a more descriptive name. And, if you select the Apply Cell Format option, the column labels will be highlighted in blue and the entire data set will have outside and inside borders. You can always select the Data Set Manager at any time in your analysis to make these types of changes.

Recommended Application Settings

StatTools allows the user to specify some of the application settings that control such things as where statistical output is displayed and how calculations are performed. The following steps show how to access the StatTools—Application Settings dialog box.

Step 1. Click the **StatTools** tab on the Ribbon
Step 2. In the **Tools Group**, click **Utilities**
Step 3. Choose **Application Settings** from the list of options

FIGURE 1.10 THE STATTOOLS—DATA SET MANAGER DIALOG BOX



StatTools - Data Set Manager [Morningstar.xlsx]

Data Set #1

Data Set

Name: Data Set #1

Excel Range: A1:F26

□ Apply Cell Formatting

Variables

Layout: ● Columns ○ Rows ☑ Names in First Row

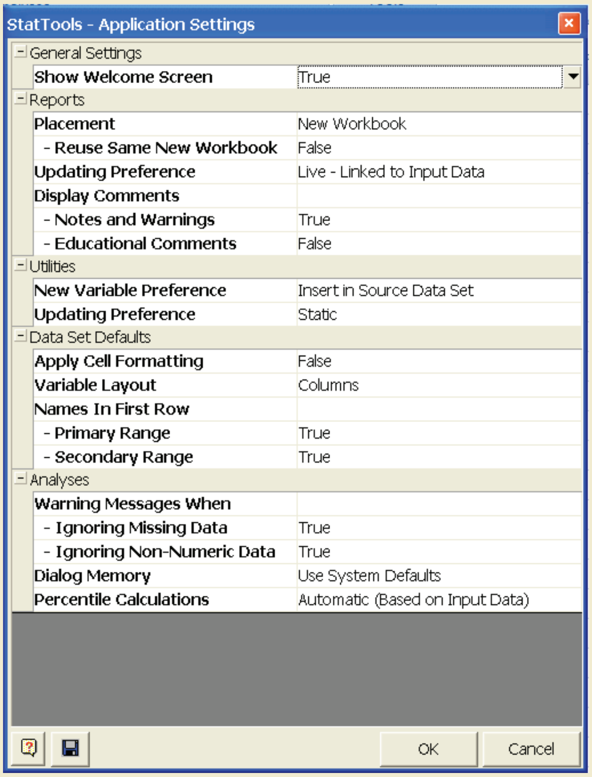
	Excel Data Range	Variable Name	Excel Range Name	Output Format
▶	A2:A26	Fund Name	auto	auto
2	B2:B26	Fund Type	auto	auto
3	C2:C26	Net Asset Value ($)	auto	auto
4	D2:D26	5 Year Average Ret.	auto	auto
5	E2:E26	Expense Ratio (%)	auto	auto
6	F2:F26	Morningstar Rank	auto	auto

6 Variables, 25 Data Cells Per Variable

Figure 1.11 shows that the StatTools—Application Settings dialog box has five sections: General Settings; Reports; Utilities; Data Set Defaults; and Analyses. Let us show how to make changes in the Reports section of the dialog box.

Figure 1.11 shows that the Placement option currently selected is **New Workbook**. Using this option, the StatTools output will be placed in a new workbook. But suppose you would like to place the StatTools output in the current (active) workbook. If you click the words **New Workbook**, a downward-pointing arrow will appear to the right. Clicking this arrow will display a list of all the placement options, including **Active Workbook**; we recommend using this option. Figure 1.11 also shows that the Updating Preferences option in the Reports section is currently **Live—Linked to Input Data**. With live updating, anytime one or more data values are changed StatTools will automatically change the output previously produced; we also recommend using this option. Note that there are two options available under Display Comments: **Notes and Warnings** and **Educational Comments**. Because these options provide useful notes and information regarding the output, we recommend using both options. Thus, to include educational

FIGURE 1.11 THE STATTOOLS—APPLICATION SETTINGS DIALOG BOX



General Settings	
Show Welcome Screen	True
Reports	
Placement	New Workbook
- Reuse Same New Workbook	False
Updating Preference	Live - Linked to Input Data
Display Comments	
- Notes and Warnings	True
- Educational Comments	False
Utilities	
New Variable Preference	Insert in Source Data Set
Updating Preference	Static
Data Set Defaults	
Apply Cell Formatting	False
Variable Layout	Columns
Names In First Row	
- Primary Range	True
- Secondary Range	True
Analyses	
Warning Messages When	
- Ignoring Missing Data	True
- Ignoring Non-Numeric Data	True
Dialog Memory	Use System Defaults
Percentile Calculations	Automatic (Based on Input Data)

comments as part of the StatTools output you will have to change the value of False for Educational Comments to True.

The StatTools—Settings dialog box contains numerous other features that enable you to customize the way that you want StatTools to operate. You can learn more about these features by selecting the Help option located in the Tools group, or by clicking the Help icon located in the lower left-hand corner of the dialog box. When you have finish making changes in the application settings, click OK at the bottom of the dialog box and then click Yes when StatTools asks you if you want to save the new application settings.