Business Statistics

A Decision-Making Approach

Groebner Shannon Fry

Ninth Edition

PEARSON NEW INTERNATIONAL EDITION

Pearson New International Edition

Business Statistics
A Decision-Making Approach
Groebner Shannon Fry
Ninth Edition

PEARSON

Pearson Education Limited
Edinburgh Gate
Harlow
Essex CM20 2JE
England and Associated Companies throughout the world

Visit us on the World Wide Web at: www.pearsoned.co.uk

PEARSON

ISBN 10: 1-292-02335-X
ISBN 13: 978-1-292-02335-9

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Printed in the United States of America

Table of Contents

1. The Where, Why, and How of Data Collection
David F. Groebner/Patrick W. Shannon/Phillip C. Fry/Kent D. Smith **1**

2. Graphs, Charts, and Tables - Describing Your Data
David F. Groebner/Patrick W. Shannon/Phillip C. Fry/Kent D. Smith **33**

3. Describing Data Using Numerical Measures
David F. Groebner/Patrick W. Shannon/Phillip C. Fry/Kent D. Smith **87**

4. Special Review Section I
David F. Groebner/Patrick W. Shannon/Phillip C. Fry/Kent D. Smith **143**

5. Introduction to Probability
David F. Groebner/Patrick W. Shannon/Phillip C. Fry/Kent D. Smith **151**

6. Discrete Probability Distributions
David F. Groebner/Patrick W. Shannon/Phillip C. Fry/Kent D. Smith **197**

7. Introduction to Continuous Probability Distributions
David F. Groebner/Patrick W. Shannon/Phillip C. Fry/Kent D. Smith **243**

8. Introduction to Sampling Distributions
David F. Groebner/Patrick W. Shannon/Phillip C. Fry/Kent D. Smith **277**

9. Estimating Single Population Parameters
David F. Groebner/Patrick W. Shannon/Phillip C. Fry/Kent D. Smith **319**

10. Introduction to Hypothesis Testing
David F. Groebner/Patrick W. Shannon/Phillip C. Fry/Kent D. Smith **363**

11. Estimation and Hypothesis Testing for Two Population Parameters
David F. Groebner/Patrick W. Shannon/Phillip C. Fry/Kent D. Smith **417**

12. Hypothesis Tests and Estimation for Population Variances
David F. Groebner/Patrick W. Shannon/Phillip C. Fry/Kent D. Smith **469**

13. Analysis of Variance
David F. Groebner/Patrick W. Shannon/Phillip C. Fry/Kent D. Smith **497**

14. Special Review Section II
David F. Groebner/Patrick W. Shannon/Phillip C. Fry/Kent D. Smith 551

15. Goodness-of-Fit Tests and Contingency Analysis
David F. Groebner/Patrick W. Shannon/Phillip C. Fry/Kent D. Smith 569

16. Introduction to Linear Regression and Correlation Analysis
David F. Groebner/Patrick W. Shannon/Phillip C. Fry/Kent D. Smith 601

17. Multiple Regression Analysis and Model Building
David F. Groebner/Patrick W. Shannon/Phillip C. Fry/Kent D. Smith 657

18. Analyzing and Forecasting Time-Series Data
David F. Groebner/Patrick W. Shannon/Phillip C. Fry/Kent D. Smith 733

19. Introduction to Nonparametric Statistics
David F. Groebner/Patrick W. Shannon/Phillip C. Fry/Kent D. Smith 797

20. Introduction to Quality and Statistical Process Control
David F. Groebner/Patrick W. Shannon/Phillip C. Fry/Kent D. Smith 831

Index 861

Quick Prep Links

- **Locate** a recent copy of a business periodical, such as *Fortune* or *Business Week*, and take note of the graphs, charts, and tables that are used in the articles and advertisements.
- **Recall** any recent experiences you have had in which you were asked to complete a written survey or respond to a telephone survey.
- **Make sure** that you have access to Excel software. Open Excel and familiarize yourself with the software.

The Where, Why, and How of Data Collection

1 What Is Business Statistics?

2 Procedures for Collecting Data ← **Outcome 1.** Know the key data collection methods.

3 Populations, Samples, and Sampling Techniques ← **Outcome 2.** Know the difference between a population and a sample.

Outcome 3. Understand the similarities and differences between different sampling methods.

4 Data Types and Data Measurement Levels ← **Outcome 4.** Understand how to categorize data by type and level of measurement.

5 A Brief Introduction to Data Mining ← **Outcome 5.** Become familiar with the concept of data mining and some of its applications.

Why you need to know

A transformation is taking place in many organizations involving how managers are using data to help improve their decision making. Because of the recent advances in software and database systems, managers are able to analyze data in more depth than ever before. A new discipline called **data mining** is growing, and one of the fastest-growing career areas is referred to as **business intelligence**. Data mining or knowledge discovery is an interdisciplinary field involving primarily computer science and statistics. People working in this field are referred to as “data scientists.” Doing an Internet search on data mining will yield a large number of sites talking about the field.

Data Mining
The application of statistical techniques and algorithms to the analysis of large data sets.

Business Intelligence
The application of tools and technologies for gathering, storing, retrieving, and analyzing data that businesses collect and use.

In today’s workplace, you can have an immediate competitive edge over other new employees, and even those with more experience, by applying statistical analysis skills to real-world decision making. The purpose of this text is to assist in your learning process and to complement your instructor’s efforts in conveying how to apply a variety of important statistical procedures.

The major automakers such as GM, Ford, and Toyota maintain databases with information on production, quality, customer satisfaction, safety records, and much more. Walmart, the world’s largest retail chain, collects and manages massive amounts of data related to the operation of its stores throughout the world. Its highly sophisticated database systems contain sales data, detailed customer data, employee satisfaction data, and much more. Governmental agencies amass extensive data on such things as unemployment, interest rates, incomes, and education. However, access to data is not limited to large companies. The relatively low cost of computer hard drives with 100-gigabyte or larger capacities makes it possible for small firms and even individuals to store vast amounts of

Anton Foltin/Shutterstock

From Chapter 1 of *Business Statistics, A Decision-Making Approach*, Ninth Edition. David F. Groebner, Patrick W. Shannon and Phillip C. Fry. Copyright © 2014 by Pearson Education, Inc. All rights reserved.

data on desktop computers. But without some way to transform the data into useful information, the data these companies have gathered are of little value.

Transforming data into information is where business statistics comes in—the statistical procedures introduced in this text are those that are used to help transform data into information. This text focuses on the practical application of statistics; we do not develop the theory you would find in a mathematical statistics course. Will you need to use math in this course? Yes, but mainly the concepts covered in your college algebra course.

Statistics does have its own terminology. You will need to learn various terms that have special statistical meaning. You will also learn certain dos and don'ts related to statistics. But most importantly, you will learn specific methods to effectively convert data into information. Don't try to memorize the concepts; rather, go to the next level of learning called *understanding*. Once you understand the underlying concepts, you will be able to *think statistically*.

Because data are the starting point for any statistical analysis, this text is devoted to discussing various aspects of data, from how to collect data to the different types of data that you will be analyzing. You need to gain an understanding of the where, why, and how of data and data collection.

1 What Is Business Statistics?

Articles in your local newspaper, news stories on television, and national publications such as the *Wall Street Journal* and *Fortune* discuss stock prices, crime rates, government-agency budgets, and company sales and profit figures. These values are statistics, but they are just a small part of the discipline called **business statistics**, which provides a wide variety of methods to assist in data analysis and decision making.

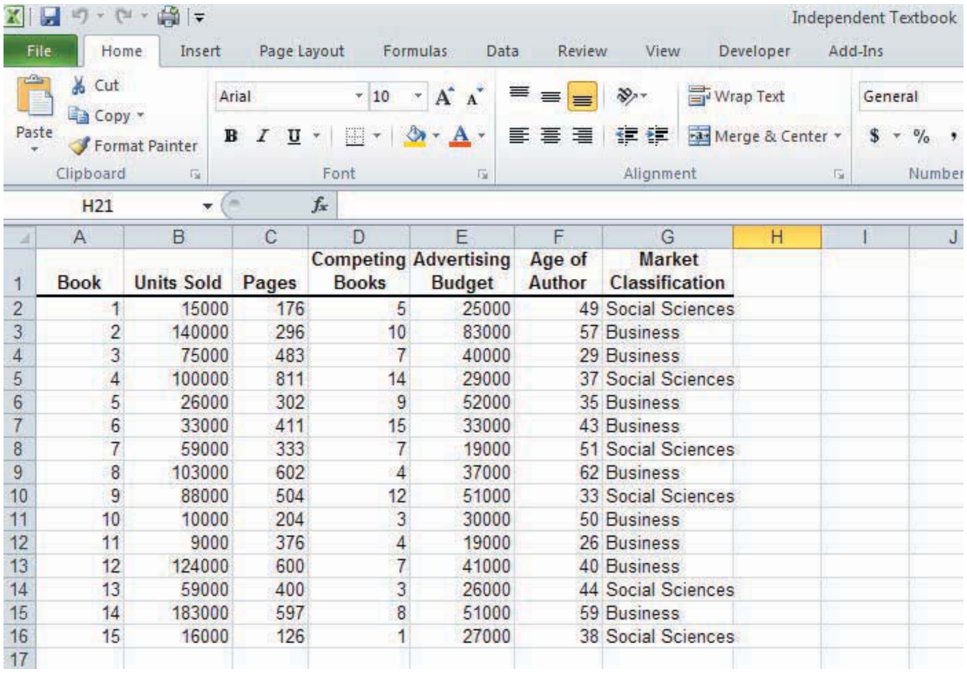
Business Statistics
A collection of procedures and techniques that are used to convert data into meaningful information in a business environment.

Descriptive Statistics

Business statistics can be segmented into two general categories. The first category involves the procedures and techniques designed to *describe data*, such as charts, graphs, and numerical measures. The second category includes tools and techniques that help decision makers *draw inferences* from a set of data. Inferential procedures include estimation and hypothesis testing. A brief discussion of these techniques follows.

BUSINESS APPLICATION DESCRIBING DATA

INDEPENDENT TEXTBOOK PUBLISHING, INC. Independent Textbook Publishing, Inc. publishes 15 college-level texts in the business and social sciences areas. Figure 1 shows an Excel spreadsheet containing data for each of these 15 textbooks. Each column



	A	B	C	D	E	F	G	H	I	J
1	Book	Units Sold	Pages	Competing Books	Advertising Budget	Age of Author	Market Classification			
2	1	15000	176	5	25000	49	Social Sciences			
3	2	140000	296	10	83000	57	Business			
4	3	75000	483	7	40000	29	Business			
5	4	100000	811	14	29000	37	Social Sciences			
6	5	26000	302	9	52000	35	Business			
7	6	33000	411	15	33000	43	Business			
8	7	59000	333	7	19000	51	Social Sciences			
9	8	103000	602	4	37000	62	Business			
10	9	88000	504	12	51000	33	Social Sciences			
11	10	10000	204	3	30000	50	Business			
12	11	9000	376	4	19000	26	Business			
13	12	124000	600	7	41000	40	Business			
14	13	59000	400	3	26000	44	Social Sciences			
15	14	183000	597	8	51000	59	Business			
16	15	16000	126	1	27000	38	Social Sciences			
17										

FIGURE 1 Excel 2010 Spreadsheet of Independent Textbook Publishing, Inc.

Excel 2010 Instructions:
1. Open File: **Independent Textbook.xlsx**.

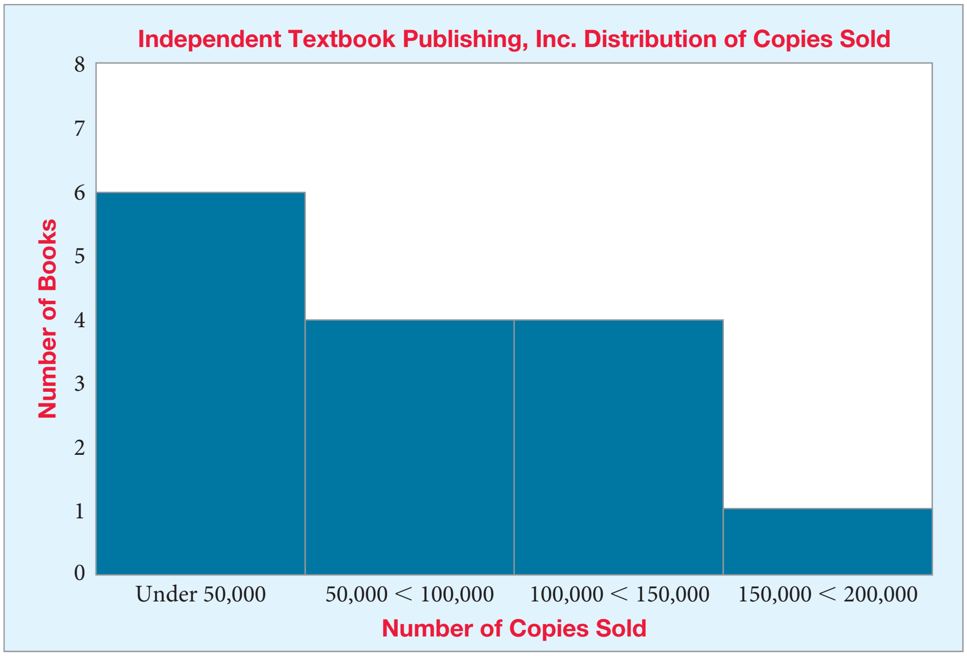


FIGURE 2 Histogram Showing the Copies Sold Distribution

in the spreadsheet corresponds to a different factor for which data were collected. Each row corresponds to a different textbook. Many statistical procedures might help the owners describe these textbook data, including descriptive techniques such as *charts*, *graphs*, and *numerical measures*.

Charts and Graphs Other text will discuss many different charts and graphs—such as the one shown in Figure 2, called a *histogram*. This graph displays the shape and spread of the distribution of number of copies sold. The *bar chart* shown in Figure 3 shows the total number of textbooks sold broken down by the two markets, business and social sciences.

Bar charts and histograms are only two of the techniques that could be used to graphically analyze the data for the textbook publisher.

BUSINESS APPLICATION DESCRIBING DATA

CROWN INVESTMENTS At Crown Investments, a senior analyst is preparing to present data to upper management on the 100 fastest-growing companies on the Hong Kong Stock Exchange. Figure 4 shows an Excel worksheet containing a subset of the data. The columns correspond to the different items of interest (growth percentage, sales, and so on). The data for each company are in a single row. The entire data are in a file called **Fast100**.

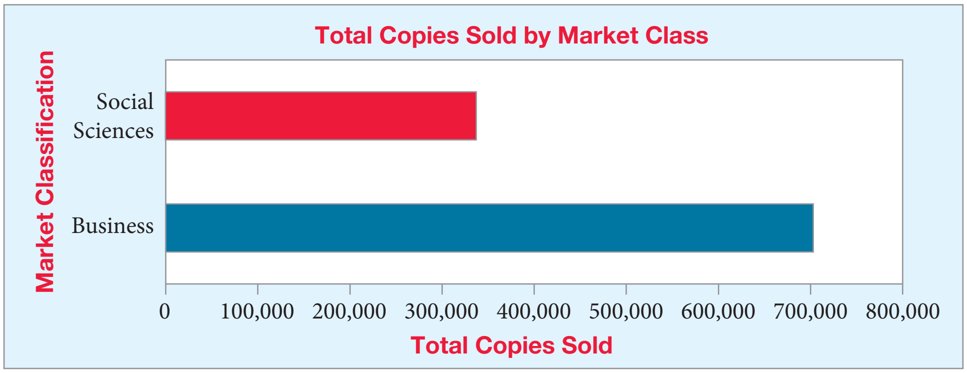
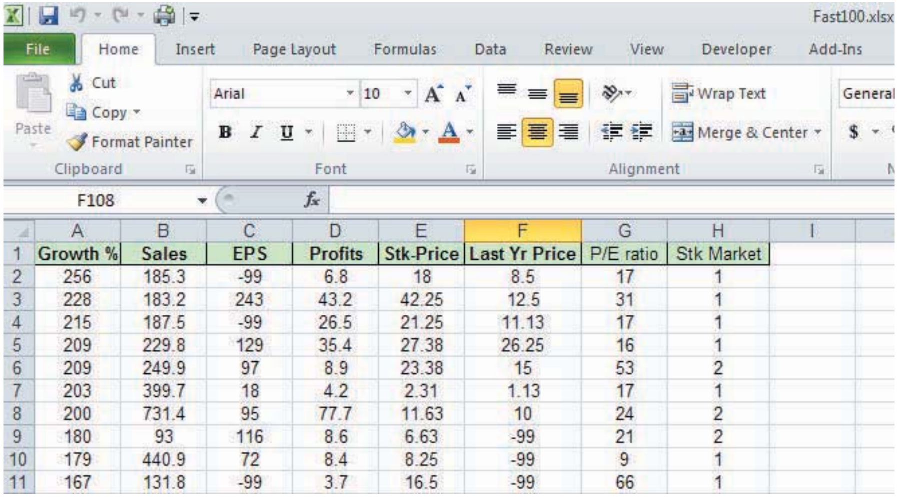


FIGURE 3 Bar Chart Showing Copies Sold by Sales Category

FIGURE 4 Crown Investment Example



	A	B	C	D	E	F	G	H	I
1	Growth %	Sales	EPS	Profits	Stk-Price	Last Yr Price	P/E ratio	Stk Market	
2	256	185.3	-99	6.8	18	8.5	17	1	
3	228	183.2	243	43.2	42.25	12.5	31	1	
4	215	187.5	-99	26.5	21.25	11.13	17	1	
5	209	229.8	129	35.4	27.38	26.25	16	1	
6	209	249.9	97	8.9	23.38	15	53	2	
7	203	399.7	18	4.2	2.31	1.13	17	1	
8	200	731.4	95	77.7	11.63	10	24	2	
9	180	93	116	8.6	6.63	-99	21	2	
10	179	440.9	72	8.4	8.25	-99	9	1	
11	167	131.8	-99	3.7	16.5	-99	66	1	

* –99 indicates missing data

Excel 2010 Instructions:
1. Open file: **Fast100.xlsx**.

Arithmetic Mean or Average
The sum of all values divided by the number of values.

In addition to preparing appropriate graphs, the analyst will compute important numerical measures. One of the most basic and most useful measures in business statistics is one with which you are already familiar: the **arithmetic mean** or **average**.

Average

The sum of all the values divided by the number of values. In equation form:

$$\text{Average} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{\text{Sum of all data values}}{\text{Number of data values}} \qquad \textbf{(1)}$$

where:

N = Number of data values
x_i = ith data value

The analyst may be interested in the average profit (that is, the average of the column labeled “Profits”) for the 100 companies. The total profit for the 100 companies is $3,193.60, but profits are given in millions of dollars, so the total profit amount is actually $3,193,600,000. The average is found by dividing this total by the number of companies:

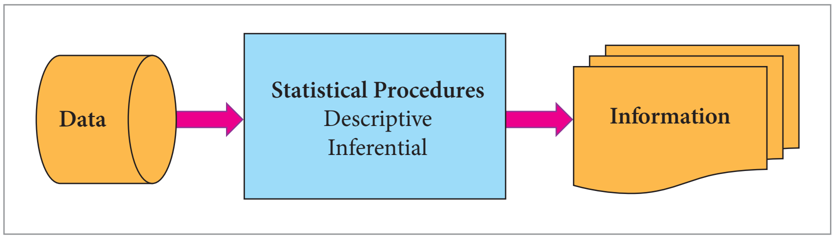
$$\text{Average} = \frac{\$3{,}193{,}600{,}000}{100} = \$31{,}936{,}000, \text{ or } \$31.936 \text{ million}$$

The average, or mean, is a measure of the center of the data. In this case, the analyst may use the average profit as an indicator—firms with above-average profits are rated higher than firms with below-average profits.

The graphical and numerical measures illustrated here are only some of the many descriptive procedures that will be introduced elsewhere. The key to remember is that the purpose of any descriptive procedure is to describe data. Your task will be to select the procedure that best accomplishes this. As Figure 5 reminds you, the role of statistics is to convert data into meaningful information.

FIGURE 5

The Role of Business Statistics



Inferential Procedures

Statistical Inference Procedures
Procedures that allow a decision maker to reach a conclusion about a set of data based on a subset of that data.

Advertisers pay for television ads based on the audience level, so knowing how many viewers watch a particular program is important; millions of dollars are at stake. Clearly, the networks don't check with everyone in the country to see if they watch a particular program. Instead, they pay a fee to the Nielsen company (http://www.nielsen.com/), which uses **statistical inference procedures** to *estimate* the number of viewers who watch a particular television program.

There are two primary categories of statistical inference procedures: *estimation* and *hypothesis testing*. These procedures are closely related but serve very different purposes.

Estimation In situations in which we would like to know about all the data in a large data set but it is impractical to work with all the data, decision makers can use techniques to estimate what the larger data set looks like. The estimates are formed by looking closely at a subset of the larger data set.

BUSINESS APPLICATION STATISTICAL INFERENCE

NEW PRODUCT INTRODUCTION Energy-boosting drinks such as Red Bull, Go Girl, Monster, and Full Throttle have become very popular among college students and young professionals. But how do the companies that make these products determine whether they will sell enough to warrant the product introduction? A typical approach is to do market research by introducing the product into one or more test markets. People in the targeted age, income, and educational categories (*target market*) are asked to sample the product and indicate the likelihood that they would purchase the product. The percentage of people who say that they will buy forms the basis for an *estimate* of the true percentage of *all* people in the target market who will buy. If that estimate is high enough, the company will introduce the product.

Hypothesis Testing Television advertising is full of product claims. For example, we might hear that "Goodyear tires will last at least 60,000 miles" or that "more doctors recommend Bayer Aspirin than any other brand." Other claims might include statements like "General Electric light bulbs last longer than any other brand" or "customers prefer McDonald's over Burger King." Are these just idle boasts, or are they based on actual data? Probably some of both! However, consumer research organizations such as Consumers Union, publisher of *Consumer Reports*, regularly test these types of claims. For example, in the hamburger case, *Consumer Reports* might select a sample of customers who would be asked to blind taste test Burger King's and McDonald's hamburgers, under the hypothesis that there is no difference in customer preferences between the two restaurants. If the sample data show a substantial difference in preferences, then the hypothesis of no difference would be rejected. If only a slight difference in preferences was detected, then *Consumer Reports* could not reject the hypothesis.

MyStatLab

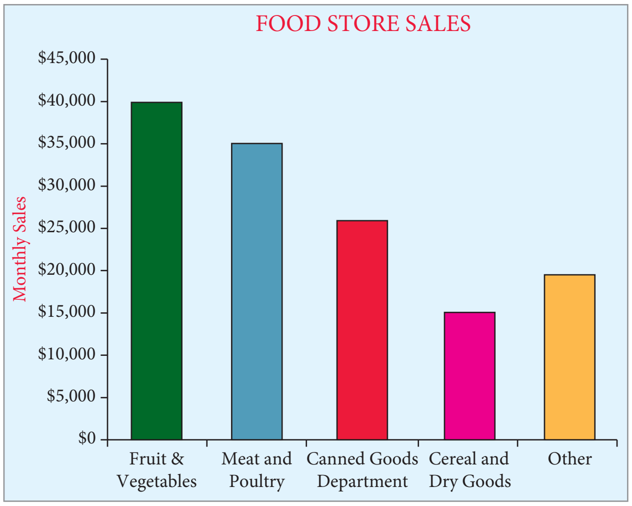
1-1: Exercises

Skill Development

1-1. For the following situation, indicate whether the statistical application is primarily descriptive or inferential.

"The manager of Anna's Fabric Shop has collected data for 10 years on the quantity of each type of dress fabric that has been sold at the store. She is interested in making a presentation that will illustrate these data effectively."

1-2. Consider the following graph that appeared in a company annual report. What type of graph is this? Explain.



1-3. Review Figures 2 and 3 and discuss any differences you see between the histogram and the bar chart.

1-4. Think of yourself as working for an advertising firm. Provide an example of how hypothesis testing can be used to evaluate a product claim.

1-5. Define what is meant by hypothesis testing. Provide an example in which you personally have tested a hypothesis (even if you didn't use formal statistical techniques to do so).

1-6. In what situations might a decision maker need to use statistical inference procedures?

1-7. Explain under what circumstances you would use hypothesis testing as opposed to an estimation procedure.

1-8. Discuss any advantages a graph showing a whole set of data has over a single measure, such as an average.

1-9. Discuss any advantages a single measure, such as an average, has over a table showing a whole set of data.

Business Applications

1-10. Describe how statistics could be used by a business to determine if the dishwasher parts it produces last longer than a competitor's brand.

1-11. Locate a business periodical such as *Fortune* or *Forbes* or a business newspaper such as *The Wall Street Journal*. Find three examples of the use of a graph to display data. For each graph,

a. Give the name, date, and page number of the periodical in which the graph appeared.
b. Describe the main point made by the graph.
c. Analyze the effectiveness of the graphs.

1-12. The human resources manager of an automotive supply store has collected the following data showing the number of employees in each of five categories by the number of days missed due to illness or injury during the past year.

Missed Days	0–2 days	3–5 days	6–8 days	8–10 days
Employees	159	67	32	10

Construct the appropriate chart for these data. Be sure to use labels and to add a title to your chart.

1-13. Suppose *Fortune* would like to determine the average age and income of its subscribers. How could statistics be of use in determining these values?

1-14. Locate an example from a business periodical or newspaper in which estimation has been used.

a. What specifically was estimated?
b. What conclusion was reached using the estimation?
c. Describe how the data were extracted and how they were used to produce the estimation.
d. Keeping in mind the goal of the estimation, discuss whether you believe that the estimation was successful and why.
e. Describe what inferences were drawn as a result of the estimation.

1-15. Locate one of the online job Web sites and pick several job listings. For each job type, discuss one or more situations in which statistical analyses would be used. Base your answer on research (Internet, business periodicals, personal interviews, etc.). Indicate whether the situations you are describing involve descriptive statistics or inferential statistics or a combination of both.

1-16. Suppose Super-Value, a major retail food company, is thinking of introducing a new product line into a market area. It is important to know the age characteristics of the people in the market area.

a. If the executives wish to calculate a number that would characterize the "center" of the age data, what statistical technique would you suggest? Explain your answer.
b. The executives need to know the percentage of people in the market area that are senior citizens. Name the basic category of statistical procedure they would use to determine this information.
c. Describe a hypothesis the executives might wish to test concerning the percentage of senior citizens in the market area.

END EXERCISES 1-1

Chapter Outcome 1. →

2 Procedures for Collecting Data

We have defined business statistics as a set of procedures that are used to transform data into information. Before you learn how to use statistical procedures, it is important that you become familiar with different types of data collection methods.

Data Collection Methods

There are many methods and procedures available for collecting data. The following are considered some of the most useful and frequently used data collection methods:

- Experiments
- Telephone surveys
- Written questionnaires and surveys
- Direct observation and personal interviews

BUSINESS APPLICATION EXPERIMENTS

FOOD PROCESSING A company often must conduct a specific experiment or set of experiments to get the data managers need to make informed decisions. For example, Lamb Weston, McCain and the J. R. Simplot Company are the primary suppliers of french fries to McDonald's in North America. At its Caldwell, Idaho, factory, the J. R. Simplot Company has a test center that, among other things, houses a mini french fry plant used to conduct experiments on its potato manufacturing process. McDonald's has strict standards on the quality of the french fries it buys. One important attribute is the color of the fries after cooking. They should be uniformly "golden brown"—not too light or too dark.

French fries are made from potatoes that are peeled, sliced into strips, blanched, partially cooked, and then freeze-dried—not a simple process. Because potatoes differ in many ways (such as sugar content and moisture), blanching time, cooking temperature, and other factors vary from batch to batch.

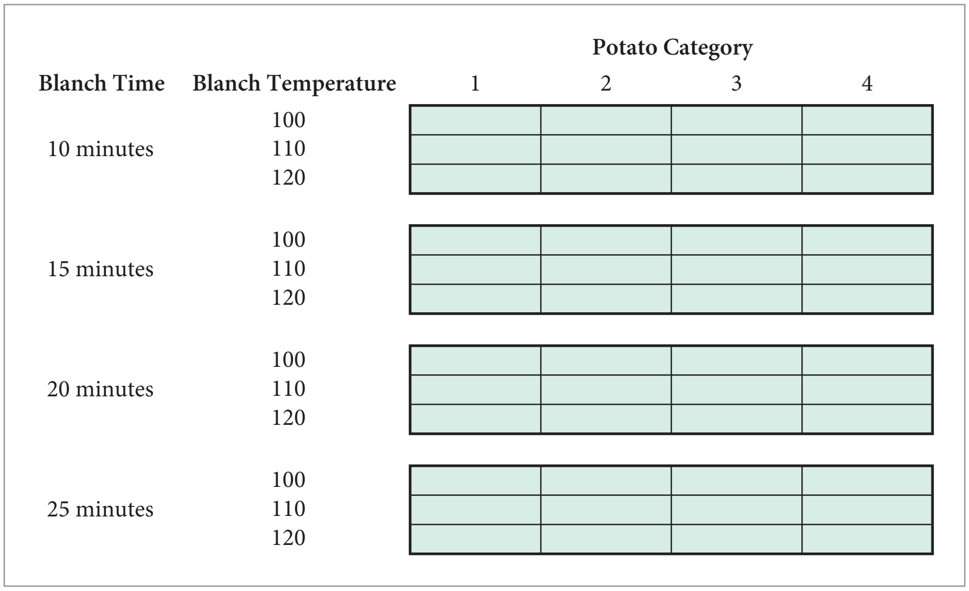
Simplot employees start their **experiments** by grouping the raw potatoes into batches with similar characteristics. They run some of the potatoes through the line with blanch time and temperature settings set at specific levels defined by an **experimental design**. After measuring one or more output variables for that run, employees change the settings and run another batch, again measuring the output variables.

Experiment
A process that produces a single outcome whose result cannot be predicted with certainty.

Experimental Design
A plan for performing an experiment in which the variable of interest is defined. One or more factors are identified to be manipulated, changed, or observed so that the impact (or influence) on the variable of interest can be measured or observed.

Figure 6 shows a typical data collection form. The output variable (for example, percentage of fries without dark spots) for each combination of potato category, blanch time, and temperature is recorded in the appropriate cell in the table.

FIGURE 6
Data Layout for the French Fry Experiment



		Potato Category			
Blanch Time	Blanch Temperature	1	2	3	4
10 minutes	100				
	110				
	120				
15 minutes	100				
	110				
	120				
20 minutes	100				
	110				
	120				
25 minutes	100				
	110				
	120				

BUSINESS APPLICATION TELEPHONE SURVEYS

PUBLIC ISSUES Chances are that you have been on the receiving end of a telephone call that begins something like: “Hello. My name is Mary Jane and I represent the XYZ organization. I am conducting a survey on …” Political groups use telephone surveys to poll people about candidates and issues. Marketing research companies use phone surveys to learn likes and dislikes of potential customers.

Telephone surveys are a relatively inexpensive and efficient data collection procedure. Of course, some people will refuse to respond to a survey, others are not home when the calls come, and some people do not have home phones—only have a cell phone—or cannot be reached by phone for one reason or another. Figure 7 shows the major steps in conducting a telephone survey. This example survey was run a few years ago by a Seattle television station to determine public support for using tax dollars to build a new football stadium for the National Football League’s Seattle Seahawks. The survey was aimed at property tax payers only.

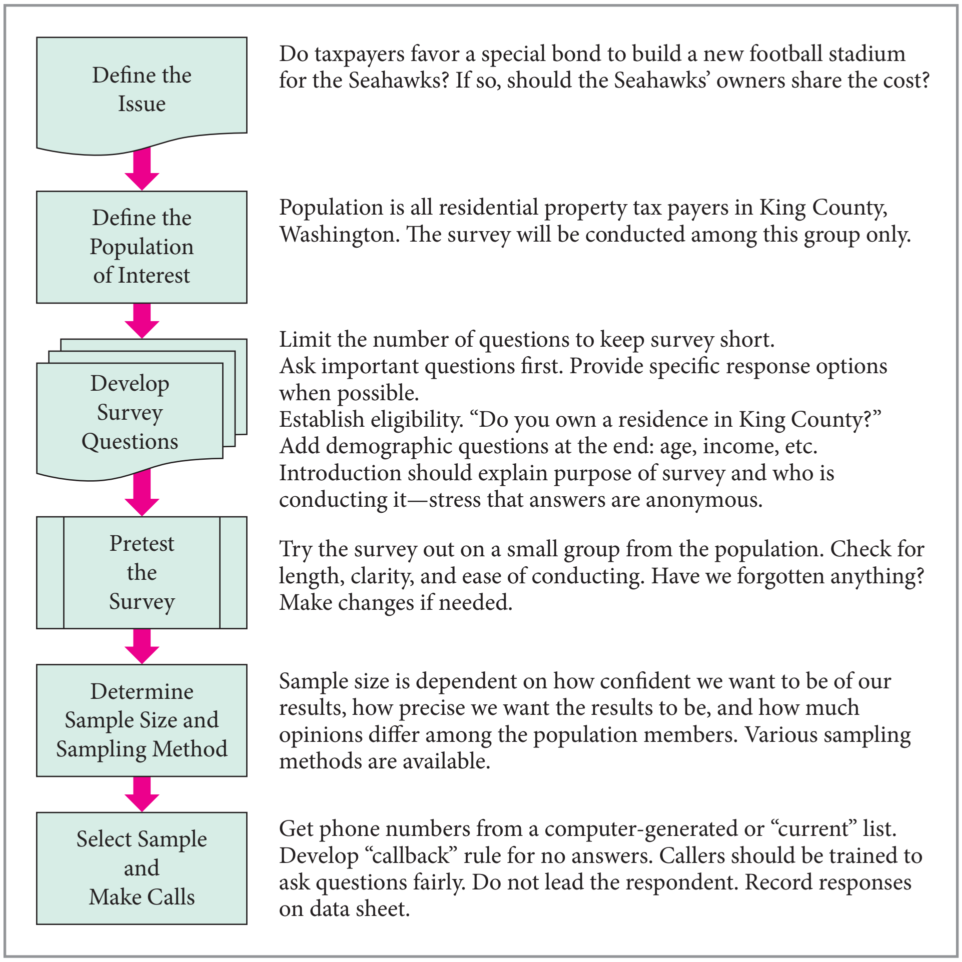
Because most people will not stay on the line very long, the phone survey must be short—usually one to three minutes. The questions are generally what are called **closed-end questions**. For example, a closed-end question might be, “To which political party do you belong? Republican? Democrat? Or other?”

Closed-End Questions
Questions that require the respondent to select from a short list of defined choices.

The survey instrument should have a short statement at the beginning explaining the purpose of the survey and reassuring the respondent that his or her responses will remain confidential. The initial section of the survey should contain questions relating to the central issue of the survey. The last part of the survey should contain **demographic questions** (such as gender, income level, education level) that will allow you to break down the responses and look deeper into the survey results.

Demographic Questions
Questions relating to the respondents’ characteristics, backgrounds, and attributes.

FIGURE 7 Major Steps for a Telephone Survey


Define the Issue
Define the Population of Interest
Develop Survey Questions
Pretest the Survey
Determine Sample Size and Sampling Method
Select Sample and Make Calls

Do taxpayers favor a special bond to build a new football stadium for the Seahawks? If so, should the Seahawks’ owners share the cost?

Population is all residential property tax payers in King County, Washington. The survey will be conducted among this group only.

Limit the number of questions to keep survey short.
Ask important questions first. Provide specific response options when possible.
Establish eligibility. “Do you own a residence in King County?”
Add demographic questions at the end: age, income, etc.
Introduction should explain purpose of survey and who is conducting it—stress that answers are anonymous.

Try the survey out on a small group from the population. Check for length, clarity, and ease of conducting. Have we forgotten anything? Make changes if needed.

Sample size is dependent on how confident we want to be of our results, how precise we want the results to be, and how much opinions differ among the population members. Various sampling methods are available.

Get phone numbers from a computer-generated or “current” list. Develop “callback” rule for no answers. Callers should be trained to ask questions fairly. Do not lead the respondent. Record responses on data sheet.

A survey budget must be considered. For example, if you have $3,000 to spend on calls and each call costs $10 to make, you obviously are limited to making 300 calls. However, keep in mind that 300 calls may not result in 300 usable responses.

The phone survey should be conducted in a short time period. Typically, the prime calling time for a voter survey is between 7:00 P.M. and 9:00 P.M. However, some people are not home in the evening and will be excluded from the survey unless there is a plan for conducting callbacks.

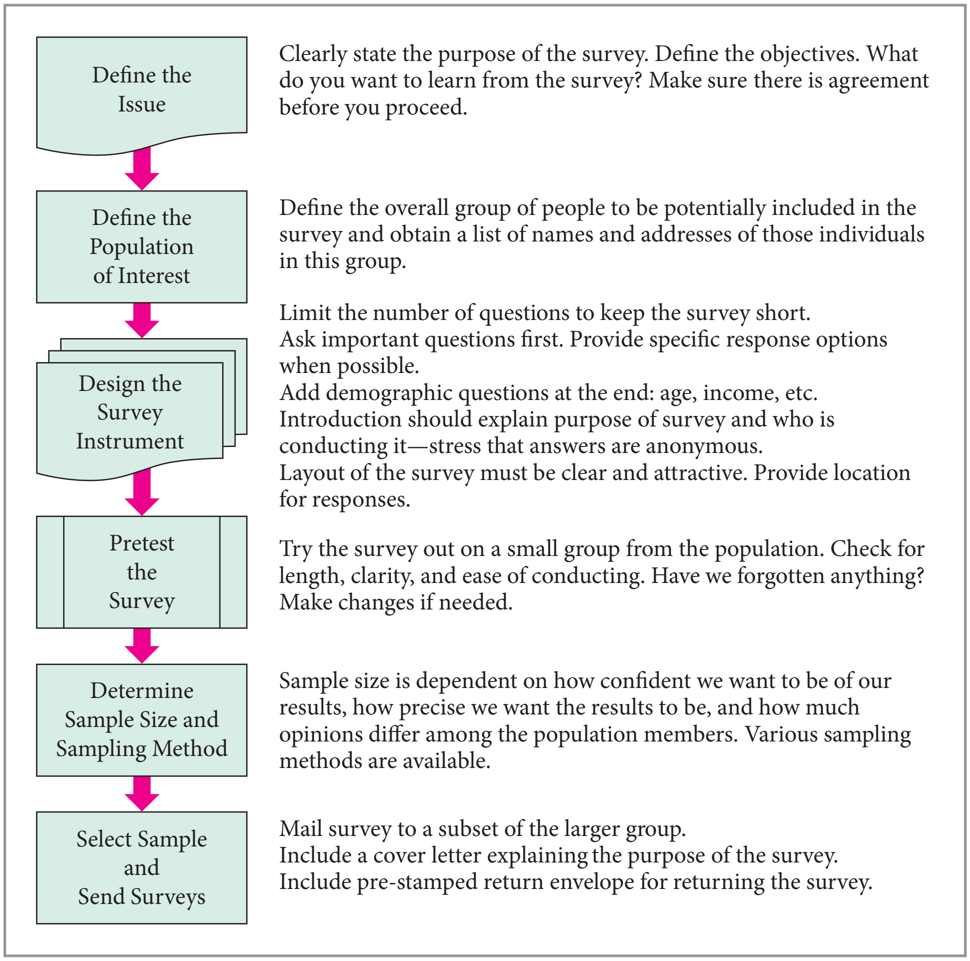
Written Questionnaires and Surveys The most frequently used method to collect opinions and factual data from people is a written questionnaire. In some instances, the questionnaires are mailed to the respondent. In others, they are administered directly to the potential respondents. Written questionnaires are generally the least expensive means of collecting survey data. If they are mailed, the major costs include postage to and from the respondents, questionnaire development and printing costs, and data analysis. Figure 8 shows the major steps in conducting a written survey. Note how written surveys are similar to telephone surveys; however, written surveys can be slightly more involved and, therefore, take more time to complete than those used for a telephone survey. However, you must be careful to construct a questionnaire that can be easily completed without requiring too much time.

Open-End Questions
Questions that allow respondents the freedom to respond with any value, words, or statements of their own choosing.

A written survey can contain both closed-end and **open-end questions**. Open-end questions provide the respondent with greater flexibility in answering a question; however, the responses can be difficult to analyze. Note that telephone surveys can use open-end questions, too. However, the caller may have to transcribe a potentially long response, and there is risk that the interviewees' comments may be misinterpreted.

Written surveys also should be formatted to make it easy for the respondent to provide accurate and reliable data. This means that proper space must be provided for the responses,

FIGURE 8 Written Survey Steps


Define the Issue
Define the Population of Interest
Design the Survey Instrument
Pretest the Survey
Determine Sample Size and Sampling Method
Select Sample and Send Surveys

Clearly state the purpose of the survey. Define the objectives. What do you want to learn from the survey? Make sure there is agreement before you proceed.

Define the overall group of people to be potentially included in the survey and obtain a list of names and addresses of those individuals in this group.

Limit the number of questions to keep the survey short.
Ask important questions first. Provide specific response options when possible.
Add demographic questions at the end: age, income, etc.
Introduction should explain purpose of survey and who is conducting it—stress that answers are anonymous.
Layout of the survey must be clear and attractive. Provide location for responses.

Try the survey out on a small group from the population. Check for length, clarity, and ease of conducting. Have we forgotten anything? Make changes if needed.

Sample size is dependent on how confident we want to be of our results, how precise we want the results to be, and how much opinions differ among the population members. Various sampling methods are available.

Mail survey to a subset of the larger group.
Include a cover letter explaining the purpose of the survey.
Include pre-stamped return envelope for returning the survey.

and the directions must be clear about how the survey is to be completed. A written survey needs to be pleasing to the eye. How it looks will affect the response rate, so it must look professional.

You also must decide whether to manually enter or scan the data gathered from your written survey. The survey design will be affected by the approach you take. If you are administering a large number of surveys, scanning is preferred. It cuts down on data entry errors and speeds up the data gathering process. However, you may be limited in the form of responses that are possible if you use scanning.

If the survey is administered directly to the desired respondents, you can expect a high response rate. For example, you probably have been on the receiving end of a written survey many times in your college career, when you were asked to fill out a course evaluation form at the end of the term. Most students will complete the form. On the other hand, if a survey is administered through the mail, you can expect a low response rate—typically 5% to 20%. Therefore, if you want 200 responses, you should mail out 1,000 to 4,000 questionnaires.

Overall, written surveys can be a low-cost, effective means of collecting data if you can overcome the problems of low response. Be careful to pretest the survey and spend extra time on the format and look of the survey instrument.

Developing a good written questionnaire or telephone survey instrument is a major challenge. Among the potential problems are the following:

- Leading questions
  - Example: "Do you agree with most other reasonably minded people that the city should spend more money on neighborhood parks?"
    - Issue: In this case, the phrase "Do you agree" may suggest that you should agree. Also, by suggesting that "most reasonably minded people" already agree, the respondent might be compelled to agree so that he or she can also be considered "reasonably minded."
    - Improvement: "In your opinion, should the city increase spending on neighborhood parks?"
  - Example: "To what extent would you support paying a small increase in your property taxes if it would allow poor and disadvantaged children to have food and shelter?"
    - Issue: The question is ripe with emotional feeling and may imply that if you don't support additional taxes, you don't care about poor children.
    - Improvement: "Should property taxes be increased to provide additional funding for social services?"
- Poorly worded questions
  - Example: "How much money do you make at your current job?"
    - Issue: The responses are likely to be inconsistent. When answering, does the respondent state the answer as an hourly figure or as a weekly or monthly total? Also, many people refuse to answer questions regarding their income.
    - Improvement: "Which of the following categories best reflects your weekly income from your current job?
      _______ Under $500 _______ $500–$1,000
      _______ Over $1,000"
  - Example: "After trying the new product, please provide a rating from 1 to 10 to indicate how you like its taste and freshness."
    - Issue: First, is a low number or a high number on the rating scale considered a positive response? Second, the respondent is being asked to rate two factors, taste and freshness, in a single rating. What if the product is fresh but does not taste good?
    - Improvement: "After trying the new product, please rate its taste on a 1 to 10 scale with 1 being best. Also rate the product's freshness using the same 1 to 10 scale.
      _______ Taste _______ Freshness"

The way a question is worded can influence the responses. Consider an example that occurred in September 2008 during the financial crisis that resulted from the sub-prime mortgage crisis and bursting of the real estate bubble. Three surveys were conducted on the same basic issue. The following questions were asked:

> "Do you approve or disapprove of the steps the Federal Reserve and Treasury Department have taken to try to deal with the current situation involving the stock market and major financial institutions?" (*ABC News/Washington Post*) 44% Approve — 42% Disapprove — 14% Unsure
>
> "Do you think the government should use taxpayers' dollars to rescue ailing private financial firms whose collapse could have adverse effects on the economy and market, or is it not the government's responsibility to bail out private companies with taxpayer dollars?" (*LA Times*/Bloomberg) 31% Use Tax Payers' Dollars — 55% Not Government's Responsibility — 14% Unsure
>
> "As you may know, the government is potentially investing billions to try and keep financial institutions and markets secure. Do you think this is the right thing or the wrong thing for the government to be doing?" (Pew Research Center) 57% Right Thing — 30% Wrong Thing—13% Unsure

Note the responses to each of these questions. The way the question is worded can affect the responses.

Direct Observation and Personal Interviews *Direct observation* is another procedure that is often used to collect data. As implied by the name, this technique requires the process from which the data are being collected to be physically observed and the data recorded based on what takes place in the process.

Possibly the most basic way to gather data on human behavior is to watch people. If you are trying to decide whether a new method of displaying your product at the supermarket will be more pleasing to customers, change a few displays and watch customers' reactions. If, as a member of a state's transportation department, you want to determine how well motorists are complying with the state's seat belt laws, place observers at key spots throughout the state to monitor people's seat belt habits. A movie producer, seeking information on whether a new movie will be a success, holds a preview showing and observes the reactions and comments of the movie patrons as they exit the screening. The major constraints when collecting observations are the amount of time and money required. For observations to be effective, trained observers must be used, which increases the cost. Personal observation is also time-consuming. Finally, personal perception is subjective. There is no guarantee that different observers will see a situation in the same way, much less report it the same way.

Structured Interview
Interviews in which the questions are scripted.

Unstructured Interview
Interviews that begin with one or more broadly stated questions, with further questions being based on the responses.

Personal interviews are often used to gather data from people. Interviews can be either **structured** or **unstructured**, depending on the objectives, and they can utilize either open-end or closed-end questions.

Regardless of the procedure used for data collection, care must be taken that the data collected are accurate and reliable and that they are the right data for the purpose at hand.

Other Data Collection Methods

Data collection methods that take advantage of new technologies are becoming more prevalent all the time. For example, many people believe that Walmart is one of the best companies in the world at collecting and using data about the buying habits of its customers. Most of the data are collected automatically as checkout clerks scan the UPC bar codes on the products customers purchase. Not only are Walmart's inventory records automatically updated, but information about the buying habits of customers is also recorded. This allows Walmart to use *analytics* and *data mining* to drill deep into the data to help with its decision making about many things, including how to organize its stores to increase sales. For instance, Walmart apparently decided to locate beer and disposable diapers close together when it discovered that many male customers also purchase beer when they are sent to the store for diapers.

Bar code scanning is used in many different data collection applications. In a DRAM (dynamic random-access memory) wafer fabrication plant, batches of silicon wafers have bar codes. As the batch travels through the plant's workstations, its progress and quality are tracked through the data that are automatically obtained by scanning.

Every time you use your credit card, data are automatically collected by the retailer and the bank. Computer information systems are developed to store the data and to provide decision makers with procedures to access the data.

In many instances, your data collection method will require you to use *physical measurement*. For example, the Andersen Window Company has quality analysts physically measure the width and height of its windows to assure that they meet customer specifications, and a state Department of Weights and Measures will physically test meat and produce scales to determine that customers are being properly charged for their purchases.

Data Collection Issues

Data Accuracy When you need data to make a decision, we suggest that you first see if appropriate data have already been collected, because it is usually faster and less expensive to use existing data than to collect data yourself. However, before you rely on data that were collected by someone else for another purpose, you need to check out the source to make sure that the data were collected and recorded properly.

Such organizations as *Bloomberg*, *Value Line*, and *Fortune* have built their reputations on providing quality data. Although data errors are occasionally encountered, they are few and far between. You really need to be concerned with data that come from sources with which you are not familiar. This is an issue for many sources on the World Wide Web. Any organization or any individual can post data to the Web. Just because the data are there doesn't mean they are accurate. Be careful.

Bias
An effect that alters a statistical result by systematically distorting it; different from a random error, which may distort on any one occasion but balances out on the average.

Interviewer Bias There are other general issues associated with data collection. One of these is the potential for **bias** in the data collection. There are many types of bias. For example, in a personal interview, the interviewer can interject bias (either accidentally or on purpose) by the way she asks the questions, by the tone of her voice, or by the way she looks at the subject being interviewed. We recently allowed ourselves to be interviewed at a trade show. The interviewer began by telling us that he would only get credit for the interview if we answered all of the questions. Next, he asked us to indicate our satisfaction with a particular display. He wasn't satisfied with our less-than-enthusiastic rating and kept asking us if we really meant what we said. He even asked us if we would consider upgrading our rating! How reliable do you think these data will be?

Nonresponse Bias Another source of bias that can be interjected into a survey data collection process is called *nonresponse bias*. We stated earlier that mail surveys suffer from a high percentage of unreturned surveys. Phone calls don't always get through, or people refuse to answer. Subjects of personal interviews may refuse to be interviewed. There is a potential problem with nonresponse. Those who respond may provide data that are quite different from the data that would be supplied by those who choose not to respond. If you aren't careful, the responses may be heavily weighted by people who feel strongly one way or another on an issue.

Selection Bias Bias can be interjected through the way subjects are selected for data collection. This is referred to as *selection bias*. A study on the virtues of increasing the student athletic fee at your university might not be best served by collecting data from students attending a football game. Sometimes, the problem is more subtle. If we do a telephone survey during the evening hours, we will miss all of the people who work nights. Do they share the same views, income, education levels, and so on as people who work days? If not, the data are biased.

Written and phone surveys and personal interviews can also yield flawed data if the interviewees *lie* in response to questions. For example, people commonly give inaccurate data about such sensitive matters as income. Lying is also an increasing problem with *exit polls* in which voters are asked who they voted for immediately after casting their vote. Sometimes, the data errors are not due to lies. The respondents may not know or have accurate information to provide the correct answer.

Observer Bias Data collection through personal observation is also subject to problems. People tend to view the same event or item differently. This is referred to as *observer bias*.

One area in which this can easily occur is in safety check programs in companies. An important part of behavioral-based safety programs is the safety observation. Trained data collectors periodically conduct a safety observation on a worker to determine what, if any, unsafe acts might be taking place. We have seen situations in which two observers will conduct an observation on the same worker at the same time, yet record different safety data. This is especially true in areas in which judgment is required on the part of the observer, such as the distance a worker is from an exposed gear mechanism. People judge distance differently.

Measurement Error A few years ago, we were working with a wood window manufacturer. The company was having a quality problem with one of its saws. A study was developed to measure the width of boards that had been cut by the saw. Two people were trained to use digital calipers and record the data. This caliper is a U-shaped tool that measures distance (in inches) to three decimal places. The caliper was placed around the board and squeezed tightly against the sides. The width was indicated on the display. Each person measured 500 boards during an 8-hour day. When the data were analyzed, it looked like the widths were coming from two different saws; one set showed considerably narrower widths than the other. Upon investigation, we learned that the person with the narrower width measurements was pressing on the calipers much more firmly. The soft wood reacted to the pressure and gave narrower readings. Fortunately, we had separated the data from the two data collectors. Had they been merged, the measurement error might have gone undetected.

Internal Validity When data are collected through experimentation, you need to make sure that proper controls have been put in place. For instance, suppose a drug company such as Pfizer is conducting tests on a drug that it hopes will reduce cholesterol. One group of test participants is given the new drug while a second group (a control group) is given a placebo. Suppose that after several months, the group using the drug saw significant cholesterol reduction. For the results to have **internal validity**, the drug company would have had to make sure the two groups were controlled for the many other factors that might affect cholesterol, such as smoking, diet, weight, gender, race, and exercise habits. Issues of internal validity are generally addressed by randomly assigning subjects to the test and control groups. However, if the extraneous factors are not controlled, there could be no assurance that the drug was the factor influencing reduced cholesterol. For data to have internal validity, the extraneous factors must be controlled.

Internal Validity
A characteristic of an experiment in which data are collected in such a way as to eliminate the effects of variables within the experimental environment that are not of interest to the researcher.

External Validity Even if experiments are internally valid, you will always need to be concerned that the results can be generalized beyond the test environment. For example, if the cholesterol drug test had been performed in Europe, would the same basic results occur for people in North America, South America, or elsewhere? For that matter, the drug company would also be interested in knowing whether the results could be replicated if other subjects are used in a similar experiment. If the results of an experiment can be replicated for groups different from the original population, then there is evidence the results of the experiment have **external validity**.

External Validity
A characteristic of an experiment whose results can be generalized beyond the test environment so that the outcomes can be replicated when the experiment is repeated.

An extensive discussion of how to measure the magnitude of bias and how to reduce bias and other data collection problems is beyond the scope of this text. However, you should be aware that data may be biased or otherwise flawed. Always pose questions about the potential for bias and determine what steps have been taken to reduce its effect.

MyStatLab

1-2: Exercises

Skill Development

1-17. If a pet store wishes to determine the level of customer satisfaction with its services, would it be appropriate to conduct an experiment? Explain.

1-18. Define what is meant by a leading question. Provide an example.

1-19. Briefly explain what is meant by an experiment and an experimental design.

1-20. Refer to the three questions discussed in this section involving the financial crises of 2008 and 2009 and possible government intervention. Note that the questions elicited different responses. Discuss the way the questions were worded and why they might have produced such different results.

1-21. Suppose a survey is conducted using a telephone survey method. The survey is conducted from 9 A.M. to 11 A.M. on Tuesday. Indicate what potential problems the data collectors might encounter.

1-22. For each of the following situations, indicate what type of data collection method you would recommend and discuss why you have made that recommendation:

a. collecting data on the percentage of bike riders who wear helmets

b. collecting data on the price of regular unleaded gasoline at gas stations in your state

c. collecting data on customer satisfaction with the service provided by a major U.S. airline

1-23. Assume you have received a class assignment to determine the attitude of students in your school toward the school's registration process. What are the validity issues you should be concerned with?

Business Applications

1-24. According to a report issued by the U.S. Department of Agriculture (USDA), the agency estimates that Southern fire ants spread at a rate of 4 to 5 miles a year. What data collection method do you think was used to collect this data? Explain your answer.

1-25. Suppose you are asked to survey students at your university to determine if they are satisfied with the food service choices on campus. What types of biases must you guard against in collecting your data?

1-26. Briefly describe how new technologies can assist businesses in their data collection efforts.

1-27. Assume you have used an online service such as Orbitz or Travelocity to make an airline reservation. The following day, you receive an e-mail containing a questionnaire asking you to rate the quality of the experience. Discuss both the advantages and disadvantages of using this form of questionnaire delivery.

1-28. In your capacity as assistant sales manager for a large office products retailer, you have been assigned the task of interviewing purchasing managers for medium and large companies in the San Francisco Bay area. The objective of the interview is to determine the office product buying plans of the company in the coming year. Develop a personal interview form that asks both issue-related questions as well as demographic questions.

1-29. The regional manager for Macy's is experimenting with two new end-of-aisle displays of the same product. An end-of-aisle display is a common method retail stores use to promote new products. You have been hired to determine which is more effective. Two measures you have decided to track are which display causes the highest percentage of people to stop and, for those who stop, which causes people to view the display the longest. Discuss how you would gather such data.

1-30. In your position as general manager for United Fitness Center, you have been asked to survey the customers of your location to determine whether they want to convert the racquetball courts to an aerobic exercise space. The plan calls for a written survey to be handed out to customers when they arrive at the fitness center. Your task is to develop a short questionnaire with at least three "issue" questions and at least three demographic questions. You also need to provide the finished layout design for the questionnaire.

1-31. According to a national CNN/USA/Gallup survey of 1,025 adults, conducted March 14–16, 2008, 63% say they have experienced a hardship because of rising gasoline prices. How do you believe the survey was conducted and what types of bias could occur in the data collection process?

END EXERCISES 1-2

Chapter Outcome 2. → 3

Populations, Samples, and Sampling Techniques

Populations and Samples

Population
The set of all objects or individuals of interest or the measurements obtained from all objects or individuals of interest.

Sample
A subset of the population.

Two of the most important terms in statistics are **population** and **sample**.

The list of all objects or individuals in the population is referred to as the *frame*. Each object or individual in the frame is known as a sampling unit. The choice of the frame depends on what objects or individuals you wish to study and on the availability of the list of these objects or individuals. Once the frame is defined, it forms the list of sampling units. The next example illustrates this concept.

BUSINESS APPLICATION **POPULATIONS AND SAMPLES**

U.S. BANK We can use U.S. Bank to illustrate the difference between a population and a sample. U.S. Bank is very concerned about the time customers spend waiting in the drive-up teller line. At a particular U.S. Bank, on a given day, 347 cars arrived at the drive-up.

A population includes measurements made on all the items of interest to the data gatherer. In our example, the U.S. Bank manager would define the population as the waiting time for all 347 cars. The list of these cars, possibly by license number, forms the frame. If she examines the entire population, she is taking a **census**. But suppose 347 cars are too many to track. The U.S. Bank manager could instead select a subset of these cars, called a *sample*. The manager could use the sample results to make statements about the population. For example, she might calculate the average waiting time for the sample of cars and then use that to conclude what the average waiting time is for the population.

Census
An enumeration of the entire set of measurements taken from the whole population.

There are trade-offs between taking a census and taking a sample. Usually the main trade-off is whether the information gathered in a census is worth the extra cost. In organizations in which data are stored on computer files, the additional time and effort of taking a census may not be substantial. However, if there are many accounts that must be manually checked, a census may be impractical.

Another consideration is that the measurement error in census data may be greater than in sample data. A person obtaining data from fewer sources tends to be more complete and thorough in both gathering and tabulating the data. As a result, with a sample there are likely to be fewer human errors.

Parameters and Statistics Descriptive numerical measures, such as an average or a proportion, that are computed from an entire population are called *parameters*. Corresponding measures for a sample are called *statistics*. Suppose in the previous example, the U.S. Bank manager timed every car that arrived at the drive-up teller on a particular day and calculated the average. This population average waiting time would be a parameter. However, if she selected a sample of cars from the population, the average waiting time for the sampled cars would be a statistic.

Sampling Techniques

Statistical Sampling Techniques
Those sampling methods that use selection techniques based on chance selection.

Nonstatistical Sampling Techniques
Those methods of selecting samples using convenience, judgment, or other nonchance processes.

Once a manager decides to gather information by sampling, he or she can use a sampling technique that falls into one of two categories: **statistical** or **nonstatistical**.

Both nonstatistical and statistical sampling techniques are commonly used by decision makers. Regardless of which technique is used, the decision maker has the same objective—to obtain a sample that is a close representative of the population. There are some advantages to using a statistical sampling technique, as we will discuss many times throughout this text. However, in many cases, nonstatistical sampling represents the only feasible way to sample, as illustrated in the following example.

BUSINESS APPLICATION **NONSTATISTICAL SAMPLING**

Elena Elisseeva/Shutterstock

SUN-CITRUS ORCHARDS Sun-Citrus Orchards owns and operates a large fruit orchard and fruit-packing plant in Florida. During harvest time in the orange grove, pickers load 20-pound sacks with oranges, which are then transported to the packing plant. At the packing plant, the oranges are graded and boxed for shipping nationally and internationally. Because of the volume of oranges involved, it is impossible to assign a quality grade to each individual orange. Instead, as each sack moves up the conveyor into the packing plant, a quality manager selects an orange sack every so often, grades the individual oranges in the sack as to size, color, and so forth, and then assigns an overall quality grade to the entire shipment from which the sample was selected.

Convenience Sampling
A sampling technique that selects the items from the population based on accessibility and ease of selection.

Because of the volume of oranges, the quality manager at Sun-Citrus uses a nonstatistical sampling method called **convenience sampling**. In doing so, the quality manager is willing to assume that orange quality (size, color, etc.) is evenly spread throughout the many sacks of oranges in the shipment. That is, the oranges in the sacks selected are of the same quality as those that were not inspected.

There are other nonstatistical sampling methods, such as *judgment sampling* and *ratio sampling*, which are not discussed here. Instead, the most frequently used statistical sampling techniques will now be discussed.

Statistical Sampling *Statistical sampling* methods (also called *probability sampling*) allow every item in the population to have a known or calculable chance of being included in the sample. The fundamental statistical sample is called a *simple random sample*. Other types of statistical sampling discussed in this text include *stratified random sampling*, *systematic sampling*, and *cluster sampling*.

Chapter Outcome 3. →

BUSINESS APPLICATION SIMPLE RANDOM SAMPLING

CABLE-ONE A salesperson at Cable-One wishes to estimate the percentage of people in a local subdivision who have satellite television service (such as Direct TV). The result would indicate the extent to which the satellite industry has made inroads into Cable-One's market. The population of interest consists of all families living in the subdivision.

For this example, we simplify the situation by saying that there are only five families in the subdivision: James, Sanchez, Lui, White, and Fitzpatrick. We will let N represent the population size and n the sample size. From the five families ($N = 5$), we select three ($n = 3$) for the sample. There are 10 possible samples of size 3 that could be selected.

{James, Sanchez, Lui}
{James, Lui, White}
{Sanchez, Lui, White}
{Lui, White, Fitzpatrick}
{James, Sanchez, White}
{James, Lui, Fitzpatrick}
{Sanchez, Lui, Fitzpatrick}
{James, Sanchez, Fitzpatrick}
{James, White, Fitzpatrick}
{Sanchez, White, Fitzpatrick}

Note that no family is selected more than once in a given sample. This method is called *sampling without replacement* and is the most commonly used method. If the families could be selected more than once, the method would be called *sampling with replacement*.

Simple Random Sampling
A method of selecting items from a population such that every possible sample of a specified size has an equal chance of being selected.

Simple random sampling is the method most people think of when they think of random sampling. In a correctly performed simple random sample, each of these samples would have an equal chance of being selected. For the Cable-One example, a simplified way of selecting a simple random sample would be to put each sample of three names on a piece of paper in a bowl and then blindly reach in and select one piece of paper. However, this method would be difficult if the number of possible samples were large. For example, if $N = 50$ and a sample of size $n = 10$ is to be selected, there are more than 10 billion possible samples. Try finding a bowl big enough to hold those!

Simple random samples can be obtained in a variety of ways. We present two examples to illustrate how simple random samples are selected in practice.

BUSINESS APPLICATION RANDOM NUMBERS

Excel
tutorials
Excel Tutorial

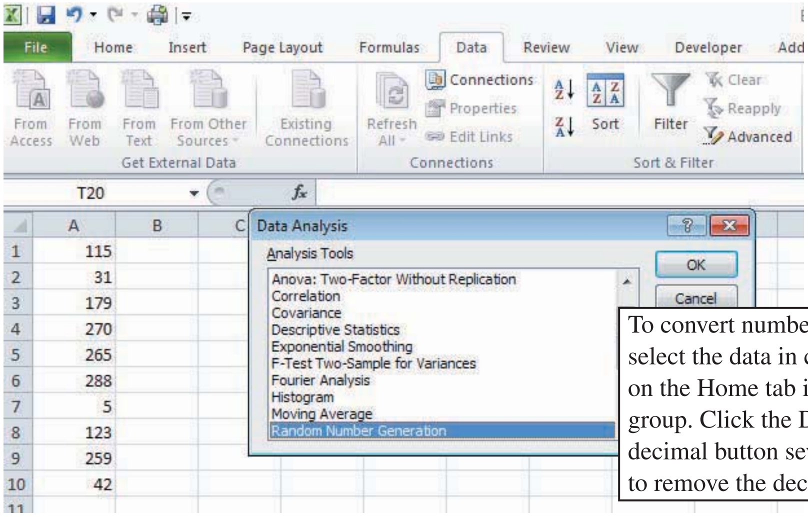
STATE SOCIAL SERVICES Suppose the state director for a Midwestern state's social services system is considering changing the timing on food stamp distribution from once a month to once every two weeks. Before making any decisions, he wants to survey a sample of 100 citizens who are on food stamps in a particular county from the 300 total food stamp recipients in that county. He first assigns recipients a number (001 to 300). He can then use the random number function in Excel to determine which recipients to include in the sample. Figure 9 shows the results when Excel chooses 10 random numbers. The first recipient sampled is number 115, followed by 31, and so forth. The important thing to remember is that assigning each recipient a number and then randomly selecting a sample from those numbers gives each possible sample an equal chance of being selected.

RANDOM NUMBERS TABLE If you don't have access to computer software such as Excel, the items in the population to be sampled can be determined by using the *random numbers table*. Begin by selecting a starting point in the random numbers table (row and digit). Suppose we use row 5, digit 8 as the starting point. Go down 5 rows and over 8 digits. Verify that the digit in this location is 1. Ignoring the blanks between columns that are there only to make the table more readable, the first three-digit number is 149. Recipient number 149 is the first one selected in the sample. Each subsequent random number is obtained from the random numbers in the next row down. For instance, the second number is 127. The procedure continues selecting numbers from top to bottom in each subsequent column. Numbers exceeding 300 and duplicate numbers are skipped. When enough numbers are

FIGURE 9 Excel 2010 Output of Random Numbers for State Social Services Example

Excel 2010 Instructions:

1. On the **Data** tab, click **Data Analysis**.
2. Select **Random Number Generation** option.
3. Set the **Number of Random Numbers** to 10.
4. Select **Uniform** as the distribution.
5. Define range as between 1 and 300.
6. Indicate that the results are to go in cell A1.
7. Click **OK**.



	A	B	C
1	115		
2	31		
3	179		
4	270		
5	265		
6	288		
7	5		
8	123		
9	259		
10	42		

To convert numbers to integers, select the data in column A and on the Home tab in the Number group. Click the Decrease decimal button several times to remove the decimal places.

found for the desired sample size, the process is completed. Food-stamp recipients whose numbers are chosen are then surveyed.

BUSINESS APPLICATION STRATIFIED RANDOM SAMPLING

Jonathan Larsen/Shutterstock

FEDERAL RESERVE BANK Sometimes, the sample size required to obtain a needed level of information from a simple random sampling may be greater than our budget permits. At other times, it may take more time to collect than is available. **Stratified random sampling** is an alternative method that has the potential to provide the desired information with a smaller sample size. The following example illustrates how stratified sampling is performed.

Stratified Random Sampling
A statistical sampling method in which the population is divided into subgroups called *strata* so that each population item belongs to only one stratum. The objective is to form strata such that the population values of interest within each stratum are as much alike as possible. Sample items are selected from each stratum using the simple random sampling method.

Each year, the Federal Reserve Board asks its staff to estimate the total cash holdings of U.S. financial institutions as of July 1. The staff must base the estimate on a sample. Note that not all financial institutions (banks, credit unions, and the like) are the same size. A majority are small, some are medium sized, and only a few are large. However, the few large institutions have a substantial percentage of the total cash on hand. To make sure that a simple random sample includes an appropriate number of small, medium, and large institutions, the sample size might have to be quite large.

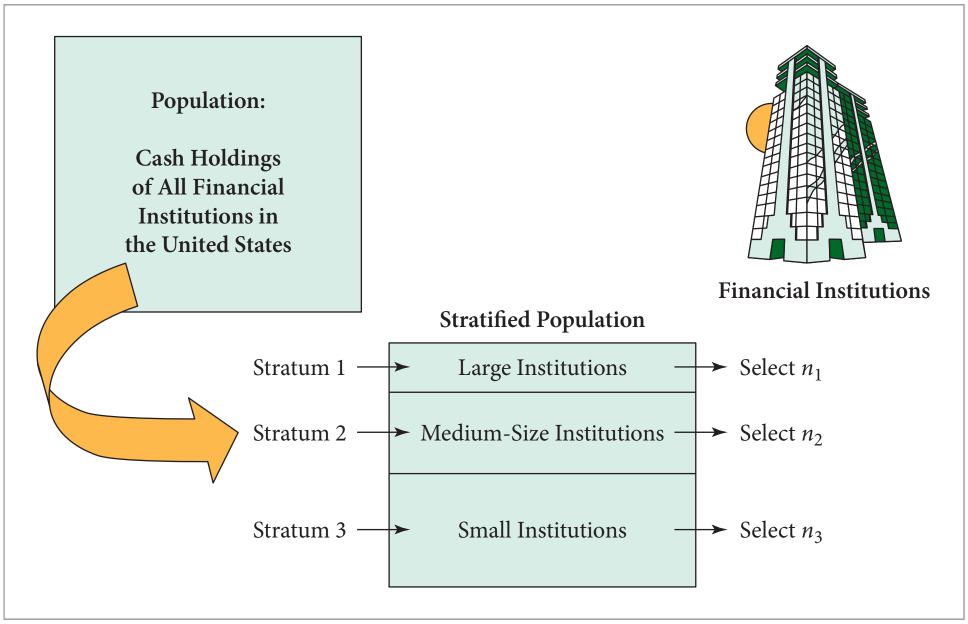
As an alternative to the simple random sample, the Federal Reserve staff could divide the institutions into three groups called *strata*: small, medium, and large. Staff members could then select a simple random sample of institutions from each stratum and estimate the total cash on hand for all institutions from this combined sample. Figure 10 shows the stratified random sampling concept. Note that the combined sample size ($n_1 + n_2 + n_3$) is the sum of the simple random samples taken from each stratum.

The key behind stratified sampling is to develop a stratum for each characteristic of interest (such as cash on hand) that has items that are quite *homogeneous*. In this example, the size of the financial institution may be a good factor to use in stratifying. Here the combined sample size ($n_1 + n_2 + n_3$) will be less than the sample size that would have been required if no stratification had occurred. Because sample size is directly related to cost (in both time and money), a stratified sample can be more cost effective than a simple random sample.

Multiple layers of stratification can further reduce the overall sample size. For example, the Federal Reserve might break the three strata in Figure 10 into *substrata* based on type of institution: state bank, interstate bank, credit union, and so on.

Most large-scale market research studies use stratified random sampling. The well-known political polls, such as the Gallup and Harris polls, use this technique also. For instance, the Gallup poll typically samples between 1,800 and 2,500 people nationwide to estimate how more than 60 million people will vote in a presidential election. We encourage you to go to the Web site http://www.gallup.com/poll/101872/how-does-gallup-polling-work.aspx to read a very good discussion about how the Gallup polls are conducted. The Web site discusses how samples are selected and many other interesting issues associated with polling.

FIGURE 10
Stratified Sampling Example



BUSINESS APPLICATION SYSTEMATIC RANDOM SAMPLING

STATE UNIVERSITY ASSOCIATED STUDENTS A few years ago, elected student council officers at mid-sized state university in the Northeast decided to survey fellow students on the issue of the legality of carrying firearms on campus. To determine the opinion of its 20,000 students, a questionnaire was sent to a sample of 500 students. Although simple random sampling could have been used, an alternative method called **systematic random sampling** was chosen.

Systematic Random Sampling
A statistical sampling technique that involves selecting every *k*th item in the population after a randomly selected starting point between 1 and *k*. The value of *k* is determined as the ratio of the population size over the desired sample size.

The university's systematic random sampling plan called for it to send the questionnaire to every 40th student (20,000/500 = 40) from an alphabetic list of all students. The process could begin by using Excel to generate a single random number in the range 1 to 40. Suppose this value was 25. The 25th student in the alphabetic list would be selected. After that, every 40th students would be selected (25, 65, 105, 145, . . .) until there were 500 students selected.

Systematic sampling is frequently used in business applications. Use it as an alternative to simple random sampling only when you can assume the population is randomly ordered with respect to the measurement being addressed in the survey. In this case, students' views on firearms on campus are likely unrelated to the spelling of their last name.

BUSINESS APPLICATION CLUSTER SAMPLING

OAKLAND RAIDERS FOOTBALL TEAM The Oakland Raiders of the National Football League plays its home games at O.co (formerly Overstock.com) Coliseum in Oakland, California. Despite its struggles to win in recent years, the team has a passionate fan base. Recently, an outside marketing group was retained by the Raiders to interview season ticket holders about the potential for changing how season ticket pricing is structured. The Oakland Raiders Web site http://www.raiders.com/tickets/seating-price-map.html shows the layout of the O.co Coliseum.

The marketing firm plans to interview season ticket holders just prior to home games during the current season. One sampling technique is to select a simple random sample of size *n* from the population of all season ticket holders. Unfortunately, this technique would likely require that interviewer(s) go to each section in the stadium. This would prove to be an expensive and time-consuming process. A systematic or stratified sampling procedure also would probably require visiting each section in the stadium. The geographical spread of those being interviewed in this case causes problems.

Cluster Sampling
A method by which the population is divided into groups, or clusters, that are each intended to be mini-populations. A simple random sample of *m* clusters is selected. The items chosen from a cluster can be selected using any probability sampling technique.

A sampling technique that overcomes the geographical spread problem is **cluster sampling**. The stadium sections would be the clusters. Ideally, the clusters would each have the same characteristics as the population as a whole.

After the clusters have been defined, a sample of *m* clusters is selected at random from the list of possible clusters. The number of clusters to select depends on various factors, including our survey budget. Suppose the marketing firm randomly selects eight clusters:

104 − 142 − 147 − 218 − 228 − 235 − 307 − 327

These are the *primary clusters*. Next, the marketing company can either survey all the ticketholders in each cluster or select a simple random sample of ticketholders from each cluster, depending on time and budget considerations.

MyStatLab

1-3: Exercises

Skill Development

1-32. Indicate which sampling method would most likely be used in each of the following situations:
a. an interview conducted with mayors of a sample of cities in Florida
b. a poll of voters regarding a referendum calling for a national value-added tax
c. a survey of customers entering a shopping mall in Minneapolis

1-33. A company has 18,000 employees. The file containing the names is ordered by employee number from 1 to 18,000. If a sample of 100 employees is to be selected from the 18,000 using systematic random sampling, within what range of employee numbers will the first employee selected come from?

1-34. Describe the difference between a statistic and a parameter.

1-35. Why is convenience sampling considered to be a nonstatistical sampling method?

1-36. Describe how systematic random sampling could be used to select a random sample of 1,000 customers who have a certificate of deposit at a commercial bank. Assume that the bank has 25,000 customers who own a certificate of deposit.

1-37. Explain why a census does not necessarily have to involve a population of people. Use an example to illustrate.

1-38. If the manager at First City Bank surveys a sample of 100 customers to determine how many miles they live from the bank, is the mean travel distance for this sample considered a parameter or a statistic? Explain.

1-39. Explain the difference between stratified random sampling and cluster sampling.

1-40. Use Excel to generate five random numbers between 1 and 900.

Business Applications

1-41. According to the U.S. Bureau of Labor Statistics, the annual percentage increase in U.S. college tuition and fees in 1995 was 6.0%, in 1999 it was 4.0%, in 2004 it was 9.5%, and in 2011 it was 5.4%. Are these percentages statistics or parameters? Explain.

1-42. According to an article in the *Idaho Statesman*, a poll taken the day before elections in Germany showed Chancellor Gerhard Schroeder behind his challenger, Angela Merkel, by 6 to 8 percentage points. Is this a statistic or a parameter? Explain.

1-43. Give the name of the kind of sampling that was most likely used in each of the following cases:
a. a *Wall Street Journal* poll of 2,000 people to determine the president's approval rating
b. a poll taken of each of the General Motors (GM) dealerships in Ohio in December to determine an estimate of the average number of Chevrolets not yet sold by GM dealerships in the United States
c. a quality-assurance procedure within a Frito-Lay manufacturing plant that tests every 1,000th bag of Fritos Corn Chips produced to make sure the bag is sealed properly
d. a sampling technique in which a random sample from each of the tax brackets is obtained by the Internal Revenue Service to audit tax returns

1-44. Your manager has given you an Excel file that contains the names of the company's 500 employees and has asked you to sample 50 employees from the list. You decide to take your sample as follows. First, you assign a random number to each employee using Excel's random number function **Rand()**. Because the random number is volatile (it recalculates itself whenever you modify the file), you freeze the random numbers using the Copy—Paste Special—Values feature. You then sort by the random numbers in ascending order. Finally, you take the first 50 sorted employees as your sample. Does this approach constitute a statistical or a nonstatistical sample?

Computer Applications

1-45. Sysco Foods is a statewide food distributor to restaurants, universities, and other establishments that prepare and sell food. The company has a very large warehouse in which the food is stored until it is pulled from the shelves to be delivered to the customers. The warehouse has 64 storage racks numbered 1−64. Each rack is three shelves high, labeled A, B, and C, and each shelf is divided into 80 sections, numbered 1−80.

Products are located by rack number, shelf letter, and section number. For example, breakfast cereal is located at 43-A-52 (rack 43, shelf A, section 52).

Each week, employees perform an inventory for a sample of products. Certain products are selected and counted. The *actual count* is compared to the *book count* (the quantity in the records that should be in stock). To simplify things, assume that the company has selected breakfast cereals to inventory. Also for simplicity's sake, suppose the cereals occupy racks 1 through 5.

a. Assume that you plan to use simple random sampling to select the sample. Use Excel to determine the sections on each of the five racks to be sampled.

b. Assume that you wish to use cluster random sampling to select the sample. Discuss the steps you would take to carry out the sampling.

c. In this case, why might cluster sampling be preferred over simple random sampling? Discuss.

1-46. United Airlines established a discount airline named Ted. The managers were interested in determining how flyers using Ted rate the airline service. They plan to question a random sample of flyers from the November 12 flights between Denver and Fort Lauderdale. A total of 578 people were on the flights that day. United has a list of the travelers together with their mailing addresses. Each traveler is given an identification number (here, from 001 to 578). Use Excel to generate a list of 40 flyer identification numbers so that those identified can be surveyed.

1-47. The National Park Service has started charging a user fee to park at selected trailheads and cross-country ski lots. Some users object to this fee, claiming they already pay taxes for these areas. The agency has decided to randomly question selected users at fee areas in Colorado to assess the level of concern.

a. Define the population of interest.

b. Assume a sample of 250 is required. Describe the technique you would use to select a sample from the population. Which sampling technique did you suggest?

c. Assume the population of users is 4,000. Use Excel to generate a list of users to be selected for the sample.

1-48. Mount Hillsdale Hospital has more than 4,000 patient files listed alphabetically in its computer system. The office manager wants to survey a statistical sample of these patients to determine how satisfied they were with service provided by the hospital. She plans to use a telephone survey of 100 patients.

a. Describe how you would attach identification numbers to the patient files; for example, how many digits (and which digits) would you use to indicate the first patient file?

b. Describe how the first random number would be obtained to begin a simple random sample method.

c. How many random digits would you need for each random number you selected?

d. Use Excel to generate the list of patients to be surveyed.

END EXERCISES 1-3

Chapter Outcome 4. → 4

Data Types and Data Measurement Levels

As you will see, the statistical techniques deal with different types of data. The level of measurement may vary greatly from application to application. In general, there are four types of data: *quantitative, qualitative, time-series,* and *cross-sectional*. A discussion of each follows.

Quantitative and Qualitative Data

Quantitative Data
Measurements whose values are inherently numerical.

Qualitative Data
Data whose measurement scale is inherently categorical.

In some cases, data values are best expressed in purely numerical, or **quantitative**, terms, such as in dollars, pounds, inches, or percentages. As an example, a cell phone provider might collect data on the number of outgoing calls placed during a month by its customers. In another case, a sports bar could collect data on the number of pitchers of beer sold weekly.

In other situations, the observation may signify only the category to which an item belongs. Categorical data are referred to as **qualitative** data.

For example, a bank might conduct a study of its outstanding real estate loans and keep track of the marital status of the loan customer—*single, married, divorced,* or *other.* The same study also might examine the credit status of the customer—*excellent, good, fair,* or *poor.* Still another part of the study might ask the customers to rate the service by the bank on a 1 to 5 scale with 1 = very poor, 2 = poor, 3 = neutral, 4 = good, and 5 = very good. Note, although the customers are asked to record a number (1 to 5) to indicate the service quality, the data would still be considered qualitative because the numbers are just codes for the categories.

Time-Series Data and Cross-Sectional Data

Time-Series Data
A set of consecutive data values observed at successive points in time.

Cross-Sectional Data
A set of data values observed at a fixed point in time.

Data may also be classified as being either **time-series** or **cross-sectional**.

The data collected by the bank about its loan customers would be cross-sectional because the data from each customer relates to a fixed point in time. In another case, if we sampled 100 stocks from the stock market and determined the closing stock price on March 15, the data would be considered cross-sectional because all measurements corresponded to one point in time.

On the other hand, Ford Motor Company tracks the sales of its F-150 pickup trucks on a monthly basis. Data values observed at intervals over time are referred to as time-series data. If we determined the closing stock price for a particular stock on a daily basis for a year, the stock prices would be time-series data.

Data Measurement Levels

Data can also be identified by their *level of measurement*. This is important because the higher the data level, the more sophisticated the analysis that can be performed.

We shall discuss and give examples of four levels of data measurements: *nominal, ordinal, interval,* and *ratio*. Figure 11 illustrates the hierarchy among these data levels, with nominal data being the lowest level.

Nominal Data *Nominal data* are the lowest form of data, yet you will encounter this type of data many times. Assigning codes to categories generates nominal data. For example, a survey question that asks for marital status provides the following responses:

1. Married   2. Single   3. Divorced   4. Other

For each person, a code of 1, 2, 3, or 4 would be recorded. These codes are nominal data. Note that the values of the code numbers have no specific meaning, because the order of the categories is arbitrary. We might have shown it this way:

1. Single   2. Divorced   3. Married   4. Other

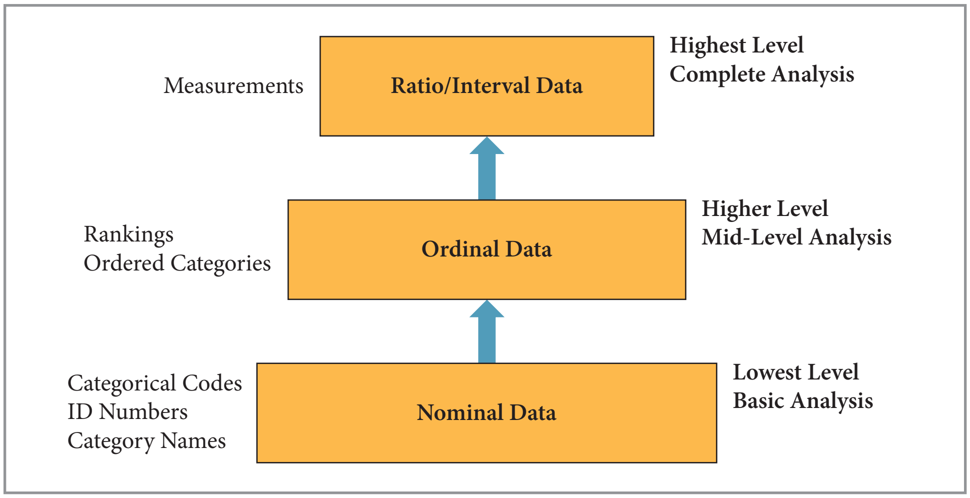
With nominal data, we also have complete control over what codes are used. For example, we could have used

88. Single   11. Divorced   33. Married   55. Other

All that matters is that you know which code stands for which category. Recognize also that the codes need not be numeric. We might use

S = Single   D = Divorced   M = Married   O = Other

FIGURE 11 Data Level Hierarchy



Ordinal Data *Ordinal* or *rank data* are one notch above nominal data on the measurement hierarchy. At this level, the data elements can be rank-ordered on the basis of some relationship among them, with the assigned values indicating this order. For example, a typical market research technique is to offer potential customers the chance to use two unidentified brands of a product. The customers are then asked to indicate which brand they prefer. The brand eventually offered to the general public depends on how often it was the preferred test brand. The fact that an ordering of items took place makes this an ordinal measure.

Bank loan applicants are asked to indicate the category corresponding to their household incomes:

________ Under $20,000	________ $20,000 to $40,000	________ over $40,000
(1)	(2)	(3)

The codes 1, 2, and 3 refer to the particular income categories, with higher codes assigned to higher incomes.

Ordinal measurement allows decision makers to equate two or more observations or to rank-order the observations. In contrast, nominal data can be compared only for equality. You cannot order nominal measurements. Thus, a primary difference between ordinal and nominal data is that ordinal data can have both an equality (=) and a greater than (>) or a less than (<) relationship, whereas nominal data can have only an equality (=) relationship.

Interval Data If the distance between two data items can be measured on some scale and the data have ordinal properties (>, <, or =) the data are said to be *interval data*. The best example of interval data is the temperature scale. Both the Fahrenheit and Celsius temperature scales have ordinal properties of ">" or "<" and "=" In addition, the distances between equally spaced points are preserved. For example, 32°F > 30°F, and 80°C > 78°C. The difference between 32°F and 30°F is the same as the difference between 80°F and 78°F, two degrees in each case. Thus, interval data allow us to precisely measure the difference between any two values. With ordinal data this is not possible, because all we can say is that one value is larger than another.

Ratio Data Data that have all the characteristics of interval data but also have a true zero point (at which zero means "none") are called *ratio data*. Ratio measurement is the highest level of measurement.

Packagers of frozen foods encounter ratio measures when they pack their products by weight. Weight, whether measured in pounds or grams, is a ratio measurement because it has a unique zero point—zero meaning no weight. Many other types of data encountered in business environments involve ratio measurements, for example, distance, money, and time.

The difference between interval and ratio measurements can be confusing because it involves the definition of a true zero. If you have $5 and your brother has $10, he has twice as much money as you. If you convert the dollars to pounds, euros, yen, or pesos, your brother will still have twice as much. If your money is lost or stolen, you have no dollars. Money has a true zero. Likewise, if you travel 100 miles today and 200 miles tomorrow, the ratio of distance traveled will be 2/1, even if you convert the distance to kilometers. If on the third day you rest, you have traveled no miles. Distance has a true zero. Conversely, if today's temperature is 35°F (1.67°C) and tomorrow's is 70°F (21.11°C), is tomorrow twice as warm as today? The answer is no. One way to see this is to convert the Fahrenheit temperature to Celsius: The ratio will no longer be 2/1 (12.64/1). Likewise, if the temperature reads 0°F (−17.59°C) this does not imply that there is no temperature. It's simply colder than 10°F (−12.22°C) Also, 0°C (32°F) is not the same temperature as 0°F. Thus, temperature, measured with either the Fahrenheit or Celsius scale (an interval-level variable), does not have a true zero.

As was mentioned earlier, a major reason for categorizing data by level and type is that the methods you can use to analyze the data are partially dependent on the level and type of data you have available.

EXAMPLE 1 CATEGORIZING DATA

Joe Gough/Shutterstock

For many years, *U.S. News and World Report* has published annual rankings based on various data collected from U.S. colleges and universities. Figure 12 shows a portion of the data in the file named **Colleges and Universities**. Each column corresponds to a different variable for which data were collected. Before doing any statistical analyses with these data, *U.S. News and World Report* employees need to determine the type and level for each of the factors. Limiting the effort to only those factors that are shown in Figure 12, this is done using the following steps:

Step 1 **Identify each factor in the data set.**

The factors (or variables) in the data set shown in Figure 12 are

College Name	State	Public (1) Private (2)	Math SAT	Verbal SAT	# appli. rec'd.	# appli. accepted.	# new stud. enrolled	# FT under-grad	# PT under-grad

Each of the 10 columns represents a different factor. Data might be missing for some colleges and universities.

Step 2 **Determine whether the data are time-series or cross-sectional.**

Because each row represents a different college or university and the data are for the same year, the data are cross-sectional. Time-series data are measured over time—say, over a period of years.

Step 3 **Determine which factors are quantitative data and which are qualitative data.**

Qualitative data are codes or numerical values that represent categories. Quantitative data are those that are purely numerical. In this case, the data for the following factors are qualitative:

College Name

State

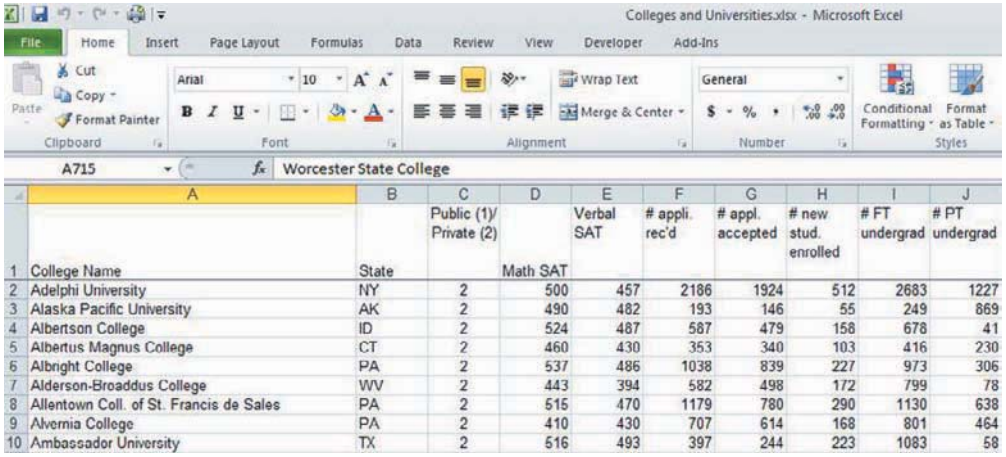
Code for Public or Private College or University

Data for the following factors are considered quantitative:

Math SAT	Verbal SAT	# new stud. enrolled
# appl. rec'd.	# appl. accepted	
# PT undergrad	# FT undergrad	

FIGURE 12

Data for U.S. Colleges and Universities



A715 *fx* Worcester State College

	A	B	C	D	E	F	G	H	I	J
1	College Name	State	Public (1)/ Private (2)	Math SAT	Verbal SAT	# appli. rec'd	# appl. accepted	# new stud. enrolled	# FT undergrad	# PT undergrad
2	Adelphi University	NY	2	500	457	2186	1924	512	2683	1227
3	Alaska Pacific University	AK	2	490	482	193	146	55	249	869
4	Albertson College	ID	2	524	487	587	479	158	678	41
5	Albertus Magnus College	CT	2	460	430	353	340	103	416	230
6	Albright College	PA	2	537	486	1038	839	227	973	306
7	Alderson-Broaddus College	WV	2	443	394	582	498	172	799	78
8	Allentown Coll. of St. Francis de Sales	PA	2	515	470	1179	780	290	1130	638
9	Alvernia College	PA	2	410	430	707	614	168	801	464
10	Ambassador University	TX	2	516	493	397	244	223	1083	58

Step 4 Determine the level of data measurement for each factor.

The four levels of data are nominal, ordinal, interval, and ratio. This data set has only nominal- and ratio-level data. The three nominal-level factors are

College Name
State
Code for Public or Private College or University

The others are all ratio-level data.

>> **END EXAMPLE**

MyStatLab

1-4: Exercises

Skill Development

1-49. For each of the following, indicate whether the data are cross-sectional or time-series:

a. quarterly unemployment rates
b. unemployment rates by state
c. monthly sales
d. employment satisfaction data for a company

1-50. What is the difference between qualitative and quantitative data?

1-51. For each of the following variables, indicate the level of data measurement:

a. product rating {1 = excellent, 2 = good, 3 = fair, 4 = poor, 5 = very poor}
b. home ownership {own, rent, other}
c. college grade point average
d. marital status {single, married, divorced, other}

1-52. What is the difference between ordinal and nominal data?

1-53. *Consumer Reports*, in its rating of cars, indicates repair history with circles. The circles are either white, black, or half and half. To which level of data does this correspond? Discuss.

Business Applications

1-54. Verizon has a support center customers can call to get questions answered about their cell phone accounts. The manager in charge of the support center has recently conducted a study in which she surveyed 2,300 customers. The customers who called the support center were transferred to a third party, who asked the customers a series of questions.

a. Indicate whether the data generated from this study will be considered cross-sectional or time-series. Explain why.
b. One of the questions asked customers was approximately how many minutes they had been on hold waiting to get through to a support person. What level of data measurement is obtained from this question? Explain.
c. Another question asked the customer to rate the service on a scale of 1–7, with 1 being the worst possible service and 7 being the best possible service. What level of data measurement is achieved from this question? Will the data be quantitative or qualitative? Explain.

1-55. The following information can be found in the Murphy Oil Corporation Annual Report to Shareholders. For each variable, indicate the level of data measurement.

a. List of Principal Offices (e.g., El Dorado, Calgary, Houston)
b. Income (in millions of dollars) from Continuing Operations
c. List of Principal Subsidiaries (e.g., Murphy Oil USA, Inc., Murphy Exploration & Production Company)
d. Number of branded retail outlets
e. Petroleum products sold, in barrels per day
f. Major Exploration and Production Areas (e.g., Malaysia, Congo, Ecuador)
g. Capital Expenditures measured in millions of dollars

1-56. You have collected the following information on 15 different real estate investment trusts (REITs). Identify whether the data are cross-sectional or time-series.

a. income distribution by region in 2012
b. per share (diluted) funds from operations (FFO) for the years 2006 to 2012
c. number of properties owned as of December 31, 2012
d. the overall percentage of leased space for the 119 properties in service as of December 31, 2012
e. dividends per share for the years 2006–2012

1-57. A loan manager for Bank of the Cascades has the responsibility for approving automobile loans. To assist her in this matter, she has compiled data on 428 cars and trucks. These data are in the file called **2004-Automobiles**.

Indicate the level of data measurement for each of the variables in this data file.

1-58. Recently, the manager of the call center for a large Internet bank asked his staff to collect data on a random sample of the bank's customers. Data on the following variables were collected and placed in a file called **Bank Call Center**:

Column A	*Column B*	*Column C*	*Column D*	*Column E*	*Column F*
Account Number	Caller Gender	Account Holder Gender	Past Due Amount	Current Amount Due	Was This a Billing Question?
Unique Tracking #	1 = Male	1 = Male	Numerical Value	Numerical Value	3 = Yes
	2 = Female	2 = Female			4 = No

A small portion of the data is as follows:

Account Number	Caller Gender	Account Holder Gender	Past Due Amount	Current Amount Due	Was This a Billing Question?
4348291	2	2	40.35	82.85	3
6008516	1	1	0	−129.67	4
17476479	1	2	0	76.38	4
13846306	2	2	0	99.24	4
21393711	1	1	0	37.98	3

a. Would you classify these data as time-series or cross-sectional? Explain.
b. Which of the variables are quantitative and which are qualitative?
c. For each of the six variables, indicate the level of data measurement.

END EXERCISES 1-4

Chapter Outcome 5. →

5 A Brief Introduction to Data Mining

Data Mining—Finding the Important, Hidden Relationships in Data

What food products have an increased demand during hurricanes? How do you win baseball games without star players? Is my best friend the one to help me find a job? What color car is least likely to be a "lemon"? These and other interesting questions can and have been answered using data mining. Data mining consists of applying sophisticated statistical techniques and algorithms to the analysis of big data (i.e., the wealth of new data that organizations collect in many and varied forms). Through the application of data mining, decisions can now be made on the basis of statistical analysis rather than on only managerial intuition and experience. The statistical techniques introduced in this text provide the basis for the more sophisticated statistical tools that are used by data mining analysts.

Wal-Mart, the nation's largest retailer, uses data mining to help it tailor product selection based on the sales, demographic, and weather information it collects. While Wal-Mart managers might not be surprised that the demand for flashlights, batteries, and bottled water increased with hurricane warnings, they were surprised to find that there was also an increase in the demand for strawberry Pop-Tarts before hurricanes hit. This knowledge allowed Wal-Mart to increase the availability of Pop-Tarts at selected stores affected by the hurricane alerts. The McKinsey Global Institute estimates that the full application of data mining to retailing could result in a potential increase in operating margins by as much as 60%. (Source: McKinsey Global Institute: ***Big Data: The Next Frontier for Innovation, Competition, and Productivity***, May 2011 by *James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers*.)