1 Defining and Collecting Data

CONTENTS

OBJECTIVES

- Understand issues that arise when defining variables
- How to define variables
- Understand the different measurement scales
- How to collect data
- Identify the different ways to collect a sample
- Understand the issues involved in data preparation
- Understand the types of survey errors

USING STATISTICS
Defining Moments

#1 You're the sales manager in charge of the best-selling beverage in its category. For years, your chief competitor has made sales gains, claiming a better tasting product. Worse, a new sibling product from your company, known for its good taste, has quickly gained significant market share at the expense of your product. Worried that your product may soon lose its number one status, you seek to improve sales by improving the product's taste. You experiment and develop a new beverage formulation. Using methods taught in this book, you conduct surveys and discover that people overwhelmingly like the newer formulation. You decide to use that new formulation going forward, having statistically shown that people prefer it. *What could go wrong?*

#2 You're a senior airline manager who has noticed that your frequent fliers always choose another airline when flying from the United States to Europe. You suspect fliers make that choice because of the other airline's perceived higher quality. You survey those fliers, using techniques taught in this book, and confirm your suspicions. You then design a new survey to collect detailed information about the quality of all components of a flight, from the seats to the meals served to the flight attendants' service. Based on the results of that survey, you approve a costly plan that will enable your airline to match the perceived quality of your competitor. *What could go wrong?*

In both cases, much did go wrong. Both cases serve as cautionary tales that if you choose the wrong variables to study, you may not end up with results that support making better decisions. Defining and collecting data, which at first glance can seem to be the simplest tasks in the DCOVA framework, can often be more challenging than people anticipate.

Coke managers also overlooked other issues, such as people's emotional connection and brand loyalty to Coca-Cola, issues better discussed in a marketing book than this book.

As the initial chapter notes, statistics is a way of thinking that can help fact-based decision making. But statistics, even properly applied using the DCOVA framework, can never be a substitute for sound management judgment. If you misidentify the business problem or lack proper insight into a problem, statistics cannot help you make a good decision. Case #1 retells the story of one of the most famous marketing blunders ever, the change in the formulation of Coca-Cola in the 1980s. In that case, Coke brand managers were so focused on the taste of Pepsi and the newly successful sibling Diet Coke that they decided only to define a variable and collect data about which drink tasters preferred in a blind taste test. When New Coke was preferred, even over Pepsi, managers rushed the new formulation into production. In doing so, those managers failed to reflect on whether the statistical results about a test that asked people to compare one-ounce samples of several beverages would demonstrate anything about beverage sales. After all, people were asked which beverage tasted better, not whether they would buy that better-tasting beverage in the future. New Coke was an immediate failure, and Coke managers reversed their decision a mere 77 days after introducing their new formulation (Polaris).

Case #2 represents a composite story of managerial actions at several airlines. In some cases, managers overlooked the need to state operational definitions for quality factors about which fliers were surveyed. In at least one case, statistics was applied correctly, and an airline spent great sums on upgrades and was able to significantly improve quality. Unfortunately, their frequent fliers still chose the competitor's flights. In this case, no statistical survey about quality could reveal the managerial oversight that given the same level of quality between two airlines, frequent fliers will almost always choose the cheaper airline. While quality was a significant variable of interest, it was not the most significant.

The lessons of these cases apply throughout this book. Due to the necessities of instruction, the examples and problems in all but the last chapter include preidentified business problems and defined variables. Identifying the business problem or objective to be considered is always a prelude to applying the DCOVA framework.

1.1 Defining Variables

Identifying a proper business problem or objective enables one to begin to identify and define the variables for analysis. For each variable identified, assign an **operational definition** that specifies the type of variable and the *scale*, the type of measurement, that the variable uses.

EXAMPLE 1.1

Defining Data at GT&M

You have been hired by Good Tunes & More (GT&M), a local electronics retailer, to assist in establishing a fair and reasonable price for Whitney Wireless, a privately held chain that GT&M seeks to acquire. You need data that would help to analyze and verify the contents of the wireless company's basic financial statements. A GT&M manager suggests that one variable you should use is monthly sales. What do you do?

SOLUTION Having first confirmed with the GT&M financial team that monthly sales is a relevant variable of interest, you develop an operational definition for this variable. Does this variable refer to sales per month for the entire chain or for individual stores? Does the variable refer to net or gross sales? Do the monthly sales data represent number of units sold or currency amounts? If the data are currency amounts, are they expressed in U.S. dollars? After getting answers to these and similar questions, you draft an operational definition for ratification by others working on this project.

Classifying Variables by Type

The type of data that a variable contains determines the statistical methods that are appropriate for a variable. Broadly, all variables are either **numerical**, variables whose data represent a counted or measured quantity, or **categorical**, variables whose data represent categories. Gender

student TIP

Some prefer the terms **quantitative** and **qualitative** over the terms numerical and categorical when describing variables. These two pairs of terms are interchangeable.

with its categories male and female is a categorical variable, as is the variable preferred-New-Coke with its categories yes and no. In Example 1.1, the monthly sales variable is numerical because the data for this variable represent a quantity.

For some statistical methods, numerical variables must be further specified as either being *discrete* or *continuous*. **Discrete** numerical variables have data that arise from a counting process. Discrete numerical variables include variables that represent a "number of something," such as the monthly number of smartphones sold in an electronics store. **Continuous** numerical variables have data that arise from a measuring process. The variable "the time spent waiting in a checkout line" is a continuous numerical variable because its data represent timing measurements. The data for a continuous variable can take on any value within a continuum or an interval, subject to the precision of the measuring instrument. For example, a waiting time could be 1 minute, 1.1 minutes, 1.11 minutes, or 1.113 minutes, depending on the precision of the electronic timing device used.

For a particular variable, one might use a numerical definition for one problem, but use a categorical definition for another problem. For example, a person's age might seem to always be a numerical age variable, but what if one was interested in comparing the buying habits of children, young adults, middle-aged persons, and retirement-age people? In that case, defining age as a categorical variable would make better sense.

Measurement Scales

Determining the **measurement scale** that the data for a variable represent is part of defining a variable. The measurement scale defines the ordering of values and determines if differences among pairs of values for a variable are equivalent and whether one value can be expressed in terms of another. Table 1.1 presents examples of measurement scales, some of which are used in the rest of this section.

TABLE 1.1
Examples of different scales and types

Data	Scale, Type	Values
Cellular provider	nominal, categorical	AT&T, T-Mobile, Verizon, Other, None
Excel skills	ordinal, categorical	novice, intermediate, expert
Temperature (°F)	interval, numerical	−459.67°F or higher
SAT Math score	interval, numerical	a value between 200 and 800, inclusive
Item cost (in $)	ratio, numerical	$0.00 or higher

Define numerical variables as using either an **interval scale**, which expresses a difference between measurements that do not include a true zero point, or a **ratio scale**, an ordered scale that includes a true zero point. Categorical variables use measurement scales that provide less insight into the values for the variable. For data measured on a **nominal scale**, category values express no order or ranking. For data measured on an **ordinal scale**, an ordering or ranking of category values is implied. Ordinal scales contain some information to compare values but not as much as interval or ratio scales. For example, the ordinal scale poor, fair, good, and excellent allows one to know that "good" is better than poor or fair and not better than excellent. But unlike interval and ratio scales, one would not know that the difference from poor to fair is the same as fair to good (or good to excellent).

PROBLEMS FOR SECTION 1.1

LEARNING THE BASICS

1.1 Four different genres of movies are playing at a movie theater: comedy, action, romance, and drama.

a. Identify whether the genre of movies is an example of a numerical or a categorical variable. Explain why.

b. Determine the measurement scale for the genres of movies shown. Explain why.

1.2 The age of a newborn baby is zero years old, which is an example of a numerical variable. Explain whether the age of a newborn baby is defined using an interval scale or a ratio scale.

1.3 The miles traveled by a company's sales representatives last month are measured.

a. Identify whether the miles traveled by sales representatives are numerical or categorical variables. Explain why.

b. Determine the measurement scale for the miles traveled by the sales representatives. Explain why.

APPLYING THE CONCEPTS

SELF TEST **1.4** For each of the following variables, determine the type of scale used and whether the variable is categorical or numerical.

a. IQ test scores
b. Car brand (Honda, BMW, Proton, or Toyota)
c. Students' performance rating scale (excellent to poor)
d. Weight (in kilogram)
e. Number of items sold per day

1.5 The following information is collected from a customer satisfaction survey conducted by a restaurant to help them drive customer loyalty and growth.

a. Gender
b. Number of food items bought
c. Number of visits to the restaurant in a week
d. Satisfaction on the variety of food and beverages available

Classify each variable as categorical or numerical and determine its measurement scale.

1.6 A university wants to link its e-learning system with one of the popular social media platforms for information-sharing activities. The different social media platforms under consideration are: Facebook, Instagram, Twitter, Skype, Blogger, and Friendster. Before selecting the platform, the university decides to analyze the students' preference through a survey. For each of the following variables, determine whether the variable is categorical or numerical and determine its measurement scale. If the variable is numerical, determine whether the variable is discrete or continuous.

a. Race of the respondent
b. The respondent's three most preferred social media platforms
c. Time, in hours, spent per day on social media
d. Most effective type of information-sharing activity on social media
e. Number of students sharing educational information on social media
f. Whether the student agreed or disagreed with the university's decision

1.7 For each of the following variables, determine whether the variable is categorical or numerical. If the variable is numerical, determine whether the variable is discrete or continuous.

a. Number of shopping trips a person made in the past month
b. A person's preferred brand of coffee
c. Amount of time a person spent exercising in the past month
d. Educational degree

1.8 Suppose the following information is collected from a teenager about his desired telecommunication package that is being offered by a local telephone company.

a. Desired telecommunication package: Postpaid 888GB
b. Monthly pocket money: $500
c. Average time spent daily on the Internet (in hours): 10.5
d. Average number of text messages sent daily: 25

Classify each of the responses by type of data and measurement scale.

1.9 A survey is conducted among a group of people who use anxiety medication. In the survey, one of the questions included is about the number of days the respondent has been using the medication. In one format, the question is "Since how many days have you been using anxiety medication?" In another format, the respondent is asked to "Tick on the circle corresponding to the number of days under anxiety medication" and is given ranges of the number of days to choose from.

a. In the first format, explain why number of days might be considered either discrete or continuous.
b. Which of these two formats would you prefer to use if you were conducting a survey? Why?

1.10 If two employees who are working in the same division both earn an income of $1,500, what arguments could be used to show that the underlying variable—ability to earn—is continuous?

1.11 Anna Johnson decides to set up an ice cream booth outside a local high school. However, Anna wants to analyze a few things before starting her business.

a. Indicate the type of data (based on the four scales: nominal, ordinal, interval, and ratio) that Anna might want to consider.
b. Identify the type of variables in (a). If the variable is numerical, determine whether the variable is discrete or continuous.

1.2 Collecting Data

Collecting data using improper methods can spoil any statistical analysis. For example, Coca-Cola managers in the 1980s (see page 47) faced advertisements from their competitor publicizing the results of a "Pepsi Challenge" in which taste testers consistently favored Pepsi over Coke. No wonder—test recruiters deliberately selected tasters they thought would likely be more favorable to Pepsi and served samples of Pepsi chilled, while serving samples of Coke lukewarm (not a very fair comparison!). These introduced biases made the challenge anything but a proper scientific or statistical test. Proper data collection avoids introducing biases and minimizes errors.

Populations and Samples

Data are collected from either a population or a sample. A **population** contains all the items or individuals of interest that one seeks to study. All of the GT&M sales transactions for a specific year, all of the full-time students enrolled in a college, and all of the registered voters in Ohio are examples of populations. A **sample** contains only a portion of a population of interest. One analyzes a sample to estimate characteristics of an entire population. For example, one might select a sample of 200 sales transactions for a retailer or select a sample of 500 registered voters in Ohio in lieu of analyzing the populations of all the sales transactions or all the registered voters.

One uses a sample when selecting a sample will be less time consuming or less cumbersome than selecting every item in the population or when analyzing a sample is less cumbersome or

learnMORE

Read the SHORT TAKES for Chapter 1 for a further discussion about data sources.

more practical than analyzing the entire population. Section FTF.3 defines *statistic* as a "value that summarizes the data of a specific variable." More precisely, a **statistic** summarizes the value of a specific variable for sample data. Correspondingly, a **parameter** summarizes the value of a population for a specific variable.

Data Sources

Data sources arise from the following activities:

- Capturing data generated by ongoing business activities
- Distributing data compiled by an organization or individual
- Compiling the responses from a survey
- Conducting an observational study and recording the results of the study
- Conducting a designed experiment and recording the outcomes of the experiment

When the person conducting an analysis performs one of these activities, the data source is a **primary data source**. When one of these activities is done by someone other than the person conducting an analysis, the data source is a **secondary data source**.

Capturing data can be done as a byproduct of, or as a result of, an organization's transactional information processing, such as the storing of sales transactions at a retailer, or as result of a service provided by a second party, such as customer information that a social media website business collects on behalf of another business. Therefore, such data capture may be either a primary or a secondary source.

Typically, organizations such as market research firms and trade associations distribute complied data, as do businesses that offer syndicated services, such as The Nielsen Company, known for its TV ratings. Therefore, this source of data is usually a secondary source. (If one supervised the distribution of a survey, compiled its results, and then analyzed those results, the survey would be a primary data source.)

Choosing to conduct an observational study or a designed experiment on a variable of interest affects the statistical methods and the decision-making processes that can be used, as Chapters 10–12 and 17 further explain.

In both observational studies and designed experiments, researchers that collect data are looking for the effect of some change, called a **treatment**, on a variable of interest. In an observational study, the researcher collects data in a natural or neutral setting and has no direct control of the treatment. For example, in an observational study of the possible effects on theme park usage patterns that a new electronic payment method might cause, one would take a sample of guests, identify those who use the new method and those who do not, and then "observe" if those who use the new method have different park usage patterns. As a designed experiment, one would select guests to use the new electronic payment method and then discover if those guests have theme park usage patterns that are different from the guests not selected to use the new payment method.

PROBLEMS FOR SECTION 1.2

APPLYING THE CONCEPTS

1.12 The quality controller at a factory that manufactures light bulbs wants to analyze the average lifetime of a light bulb. A sample of 1,000 light bulbs is tested and the average lifetime of a bulb in this sample is found to be 555 hours.

a. Identify the population and sample for the abovementioned analysis.

b. Justify whether the analysis is based on a primary or a secondary data source.

1.13 The possible effects of vitamin C and vitamin E on health is being studied. Vitamin C is taken in three different amounts of 100 mg, 250 mg, and 500 mg daily. At the same time, vitamin E can be taken either in 150 mg or 400 mg doses daily. How many different treatments are possible? List all the treatments.

1.14 Visit the official website of Bureau of Economic Analysis of United States government at https://apps.bea.gov/iTable/iTable.cfm?ReqID=62&step=1#reqid=62&step=9&isuri=1&6210=4. Retrieve the necessary data for an analysis of the U.S. international trade in goods and services. What type of data source is your analysis based on?

1.15 A study is conducted to analyze consumers' satisfaction on Anna Sui perfumes sold at an outlet store. For the study, 500 customers who purchased any Anna Sui perfumes were interviewed. The result showed that 80% of the brand's customers are satisfied with their purchase. What type of data source has been used in the study?

1.16 A team of researchers is interested in analyzing the number of cars produced in Malaysia during the first six months of the year 2019. They visit the following website: https://tradingeconomics.com/malaysia/car-production to retrieve the necessary data for this analysis. What type of data source have they based their analysis on?

1.3 Types of Sampling Methods

When selecting a sample to collect data, begin by defining the **frame**. The frame is a complete or partial listing of the items that make up the population from which the sample will be selected. Inaccurate or biased results can occur if a frame excludes certain groups, or portions of the population. Using different frames to collect data can lead to different, even opposite, conclusions.

Using the frame, select either a nonprobability sample or a probability sample. In a **nonprobability sample**, select the items or individuals without knowing their probabilities of selection. In a **probability sample**, select items based on known probabilities. Whenever possible, use a probability sample because such a sample will allow one to make inferences about the population being analyzed.

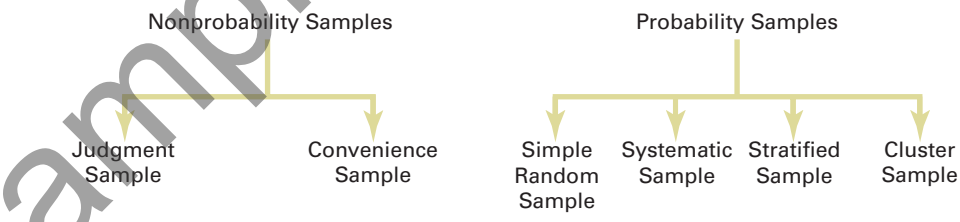
Nonprobability samples can have certain advantages, such as convenience, speed, and low cost. Such samples are typically used to obtain informal approximations or as small-scale initial or pilot analyses. However, because the theory of statistical inference depends on probability sampling, nonprobability samples *cannot be used* for statistical inference and this more than offsets those advantages in more formal analyses.

Figure 1.1 shows the subcategories of the two types of sampling. A nonprobability sample can be either a convenience sample or a judgment sample. To collect a **convenience sample**, select items that are easy, inexpensive, or convenient to sample. For example, in a warehouse of stacked items, selecting only the items located on the top of each stack and within easy reach would create a convenience sample. So, too, would be the responses to surveys that the websites of many companies offer visitors. While such surveys can provide large amounts of data quickly and inexpensively, the convenience samples selected from these responses will consist of self-selected website visitors. (Read the *Consider This* essay on page 59 for a related story.)

To collect a **judgment sample**, collect the opinions of preselected experts in the subject matter. Although the experts may be well informed, one cannot generalize their results to the population.

The types of probability samples most commonly used include simple random, systematic, stratified, and cluster samples. These four types of probability samples vary in terms of cost, accuracy, and complexity, and they are the subject of the rest of this section.

FIGURE 1.1
Types of samples



Simple Random Sample

In a **simple random sample**, every item from a frame has the same chance of selection as every other item, and every sample of a fixed size has the same chance of selection as every other sample of that size. Simple random sampling is the most elementary random sampling technique. It forms the basis for the other random sampling techniques. However, simple random sampling has its disadvantages. Its results are often subject to more variation than other sampling methods. In addition, when the frame used is very large, carrying out a simple random sample may be time consuming and expensive.

With simple random sampling, use n to represent the sample size and N to represent the frame size. Number every item in the frame from 1 to N. The chance that any particular member of the frame will be selected during the first selection is $1/N$.

Select samples with replacement or without replacement. **Sampling with replacement** means that selected items are returned to the frame, where it has the same probability of being selected again. For example, imagine a fishbowl containing N business cards, one card for each person. The first selection selects the card for Grace Kim. After her information has been recorded, her business card is placed back in the fishbowl. All cards are thoroughly mixed and a second selection is made. For this second selection, the probability that the card for Grace Kim will be selected remains $1/N$.

Most sampling is *sampling without replacement*. **Sampling without replacement** means that once an item has been selected, the item cannot ever again be selected for the sample.

The chance that any particular item in the frame will be selected—for example, the business card for Grace Kim—on the first selection is $1/N$. The chance that any card not previously chosen will be chosen on the second selection becomes 1 out of $N - 1$.

When creating a simple random sample, avoid the "fishbowl" method of selecting a sample because this method lacks the ability to thoroughly mix items and, therefore, randomly select a sample. Instead, use a more rigorous selection method.

learnMORE
Learn to use a table of random numbers to select a simple random sample in the **Section 1.3 LearnMore** online topic.

One such method is to use a **table of random numbers**, such as Table E.1 in Appendix E, for selecting the sample. A table of random numbers consists of a series of digits listed in a randomly generated sequence. To use a random number table for selecting a sample, assign code numbers to the individual items of the frame. Then generate the random sample by reading the table of random numbers and selecting those individuals from the frame whose assigned code numbers match the digits found in the table. Because every digit or sequence of digits in the table is random, the table can be read either horizontally or vertically. The margins of the table designate row numbers and column numbers, and the digits are grouped into sequences of five in order to make reading the table easier.

Because the number system uses 10 digits (0, 1, 2, … , 9), the chance that any particular digit will be randomly generated is equal 1 out of 10 and is equal to the probability of generating any other digit. For a generated sequence of 800 digits, one would expect about 80 to be the digit 0, 80 to be the digit 1, and so on.

Systematic Sample

In a **systematic sample**, partition the N items in the frame into n groups of k items, where

$$k = \frac{N}{n}$$

Round k to the nearest integer. To select a systematic sample, choose the first item to be selected at random from the first k items in the frame. Then, select the remaining $n - 1$ items by taking every kth item thereafter from the entire frame.

If the frame consists of a list of prenumbered checks, sales receipts, or invoices, taking a systematic sample is faster and easier than taking a simple random sample. A systematic sample is also a convenient mechanism for collecting data from membership directories, electoral registers, class rosters, and consecutive items coming off an assembly line.

To take a systematic sample of $n = 40$ from the population of $N = 800$ full-time employees, partition the frame of 800 into 40 groups, each of which contains 20 employees. Then select a random number from the first 20 individuals and include every twentieth individual after the first selection in the sample. For example, if the first random number selected is 008, subsequent selections will be 028, 048, 068, 088, 108, … , 768, and 788.

Simple random sampling and systematic sampling are simpler than other, more sophisticated, probability sampling methods, but they generally require a larger sample size. In addition, systematic sampling is prone to selection bias that can occur when there is a pattern in the frame. To overcome the inefficiency of simple random sampling and the potential selection bias involved with systematic sampling, one can use either stratified sampling methods or cluster sampling methods.

Stratified Sample

learnMORE
Learn how to select a stratified sample in the **Section 1.3 LearnMore** online topic.

In a **stratified sample**, first subdivide the N items in the frame into separate subpopulations, or **strata**. A stratum is defined by some common characteristic, such as gender or year in school. Then select a simple random sample within each of the strata and combine the results from the separate simple random samples. Stratified sampling is more efficient than either simple random sampling or systematic sampling because the representation of items across the entire population is assured. The homogeneity of items within each stratum provides greater precision in the estimates of underlying population parameters. In addition, stratified sampling enables one to reach conclusions about each strata in the frame. However, using a stratified sample requires that one can determine the variable(s) on which to base the stratification and can also be expensive to implement.

Cluster Sample

In a **cluster sample**, divide the N items in the frame into clusters that contain several items. **Clusters** are often naturally occurring groups, such as counties, election districts, city blocks,

households, or sales territories. Then take a random sample of one or more clusters and study all items in each selected cluster.

Cluster sampling is often more cost-effective than simple random sampling, particularly if the population is spread over a wide geographic region. However, cluster sampling often requires a larger sample size to produce results as precise as those from simple random sampling or stratified sampling. The Cochran, Groves et al., and Lohr sources discuss systematic sampling, stratified sampling, and cluster sampling procedures.

PROBLEMS FOR SECTION 1.3

LEARNING THE BASICS

1.17 For a population containing $N = 902$ individuals, what code number would you assign for

a. the ninth person on the list?
b. the twentieth person on the list?
c. the last person on the list?

1.18 For a population of $N = 950$, verify that by starting in row 3, column 1 of the table of random numbers (Table E.1), you need only five rows to select a sample of $n = 50$ *without* replacement.

1.19 Given a population of $N = 60$, starting in row 16, column 1 of the table of random numbers (Table E.1), and reading across the row, select a sample of $n = 10$

a. *without* replacement.
b. *with* replacement.

APPLYING THE CONCEPTS

1.20 The head of the student association at Taylor's University, Malaysia, would like to know what students think of the university website. Since he is unable to get a list of all students enrolled at the university, he and other members of the association stand outside the university's student center, requesting passers-by to answer a questionnaire. What type of sample and sampling method have been used?

1.21 The principal of a school in Mumbai, India, wants to know each students' favorite subject. The first-grade students will have different subject preferences than the sixth-graders. Which sampling method should the principal use to conduct an analysis that will deliver precise results?

1.22 The manager at a supermarket needs to select 10 out of 55 staff members to attend a professional training seminar. She must be unbiased in her selection. What type of sampling should the manager do? Explain how.

1.23 The registrar of a university with a population of $N = 4{,}200$ full-time students is asked by the president to conduct a survey to measure students' satisfaction with the quality of life on campus. The following table contains a breakdown of the 4,200 registered full-time students, by gender and class designation:

	CLASS DESIGNATION				
GENDER	**Fr.**	**So.**	**Jr.**	**Sr.**	**Total**
Female	507	514	563	467	2,051
Male	553	547	484	565	2,149
Total	1,060	1,061	1,047	1,032	4,200

The registrar intends to take a probability sample of $n = 200$ students and project the results from the sample to the entire population of full-time students.

a. If the frame available from the registrar's files is an alphabetical listing of the names of all $N = 4{,}200$ registered students, what types of samples could you take? Discuss.
b. What is the advantage of selecting a simple random sample in (a)?
c. What is the advantage of selecting a systematic sample in (a)?
d. If the frame available from the registrar's files is a listing of the names of all $N = 4{,}200$ registered students compiled from eight separate alphabetical lists, based on the gender and class designation breakdowns shown in the class designation table, what type of sample should you take?
e. Suppose that each of the $N = 4{,}200$ registered students lived in one of the 10 campus dormitories. Each dormitory accommodates 420 students. It is college policy to fully integrate students by gender and class designation in each dormitory. If the registrar is able to compile a listing of all students by dormitory, explain how you would take a cluster sample.

SELF TEST **1.24** The owner of an electronic store wants to conduct a survey to measure customer satisfaction for four different brands of washing machine purchased from his store over the past 12 months. His records indicate that 35 customers purchased brand A, 25 purchased brand B, 17 purchased brand C, and 23 purchased brand D.

a. If the owner decides to have a random sample of 20 customers, how many should be selected for each brand?
b. Starting in row 18, column 01, and proceeding horizontally in the table of random numbers (Table E.1), select a sample of $n = 20$ customers.
c. Name the sampling method that has been applied in (a) and (b). What is one of the advantages of selecting that method?

1.25 The Dean of Students at a university mailed out a survey to 400 students. The sample included 100 students randomly selected from each of the freshman, sophomore, junior, and senior classes on campus.

a. What type of sampling did the dean use?
b. Explain why the sampling method in (a) is the most efficient method.
c. Explain how you would carry out the sampling according to the method stated in (a).

1.4 Data Cleaning

Even if proper data collection procedures are followed, the collected data may contain incorrect or inconsistent data that could affect statistical results. **Data cleaning** corrects such defects and ensures the data contain suitable *quality* for analysis. Cleaning is the most important data preprocessing task and *must* be done before performing any analysis. Cleaning can take a significant amount of time to do. One survey of big data analysts reported that they spend 60% of their time cleaning data, while only 20% of their time collecting data and a similar percentage for analyzing data (Press).

Data cleaning seeks to correct the following types of irregularities:

- Invalid variable values, including non-numerical data for a numerical variable, invalid categorical values of a categorical variable, and numeric values outside a defined range
- Coding errors, including inconsistent categorical values, inconsistent case for categorical values, and extraneous characters
- Data integration errors, including redundant columns, duplicated rows, differing column lengths, and different units of measure or scale for numerical variables

With the exception of several examples designed for use with this section, data for the problems and examples in this book have already been properly cleaned to allow focus on the statistical concepts and methods that the book discusses.

By its nature, data cleaning cannot be a fully automated process, even in large business systems that contain data cleaning software components. As this chapter's software guides explain, Excel and Tableau contain functionality that lessens the burden of data cleaning (see the Excel and Tableau Guides for this chapter). When performing data cleaning, first preserve a copy of the original data for later reference.

Invalid Variable Values

Invalid variable values can be identified as being incorrect by simple scanning techniques so long as operational definitions for the variables the data represent exist. For any numerical variable, any value that is not a number is clearly an incorrect value. For a categorical variable, a value that does not match any of the predefined categories of the variable is, likewise, clearly an incorrect value. And for numerical variables defined with an explicit range of values, a value outside that range is clearly an error.

Coding Errors

Coding errors can result from poor recording or entry of data values or as the result of computerized operations such as copy-and-paste or data import. While coding errors are literally invalid values, coding errors may be correctable without consulting additional information whereas the invalid variable values *never* are. For example, for a Gender variable with the defined values F and M, the value "Female" is a *coding error* that can be reasonably changed to F. However, the value "New York" for the same variable is an *invalid variable value* that you cannot reasonably change to either F or M.

Unlike invalid variable values, coding errors may be *tolerated* by analysis software. For example, for the same Gender variable, the values M and m might be treated as the "same" value for purposes of an analysis by software that was tolerant of case inconsistencies, an attribute known as being *insensitive* to case.

Perhaps the most frustrating coding errors are extraneous characters in a value. Visual examination may not be able to spot extraneous characters such as nonprinting characters or extra, trailing space characters as one scans data. For example, the value David and the value that is David followed by three space characters may look the same to one casually scanning them but may not be treated the same by software. Likewise, values with nonprinting characters may look correct but may cause software errors or be reported as invalid by analysis software.

Data Integration Errors

Data integration errors arise when data from two different computerized sources, such as two different data repositories are combined into one data set for analysis. Identifying data integration errors may be the most time-consuming data cleaning task. Because spotting these errors requires a type of data interpretation that automated processes of a typical business computer

Perhaps not surprising, supplying business systems with automated data interpretation skills is a goal of many companies that provide data analysis software and services.

systems today cannot supply, spotting these errors using manual methods will be typical for the foreseeable future.

Some data integration errors occur because variable names or definitions for the same item of interest have minor differences across systems. In one system, a customer ID number may be known as Customer ID, whereas in a different system, the same variable is known as Cust Number. A result of combining data from the two systems may result in having both Customer ID and Cust Number variable columns, a redundancy that should be eliminated.

Duplicated rows also occur because of similar inconsistencies across systems. Consider a Customer Name variable with the value that represents the first coauthor of this book, David M. Levine. In one system, this name may have been recorded as David Levine, whereas in another system, the name was recorded as D M Levine. Combining records from both systems may result in two records, where only one should exist. Whether "David Levine" is actually the same person as "D M Levine" requires an interpretation skill that today's software may lack.

Likewise, different units of measurement (or scale) may not be obvious without additional, human interpretation. Consider the variable Air Temperature, recorded in degrees Celsius in one system and degrees Fahrenheit in another. The value 30 would be a plausible value under either measurement system and without further knowledge or context impossible to spot as a Celsius measurement in a column of otherwise Fahrenheit measurements.

Missing Values

Missing values are values that were not collected for a variable. For example, survey data may include answers for which no response was given by the survey taker. Such "no responses" are examples of missing values. Missing values can also result from integrating two data sources that do not have a row-to-row correspondence for each row in both sources. The lack of correspondence creates particular variable columns to be longer, to contain additional rows than the other columns. For these additional rows, *missing* would be the value for the cells in the shorter columns.

Do not confuse missing values with miscoded values. *Unresolved* miscoded values—values that cannot be cleaned by any method—might be changed to *missing* by some researchers or excluded for analysis by others.

Algorithmic Cleaning of Extreme Numerical Values

For numerical variables without a defined range of possible values, one might find **outliers**, values that seem excessively different from most of the other values. Such values may or may not be errors, but all outliers require review. While there is no one standard for defining outliers, most define outliers in terms of descriptive measures such as the standard deviation or the interquartile range that Chapter 3 discusses. Because software can compute such measures, spotting outliers can be automated if a definition of the term that uses such a measure is used. As later chapters note when appropriate, identifying outliers is important as some methods are *sensitive* to outliers and produce very different results when outliers are included in analysis.

1.5 Other Data Preprocessing Tasks

In addition to data cleaning, one might undertake several other data processing tasks before visualizing and analyzing a set of data.

Data Formatting

Data formatting includes rearranging the structure of the data or changing the electronic encoding of the data or both. For example, consider financial data that has been collected for a sample of companies. The collected data may be structured as tables of data, as the contents of standard forms, in a continuous stock ticker stream, or as messages or blog entries that appear on various websites. These data sources have various levels of structure that affect the ease of reformatting them for use.

Because tables of data are highly structured and are similar to the structure of a worksheet, tables would require the least reformatting. In the best case, the rows and columns of a table would become the rows and columns of a worksheet. Unstructured data sources, such as messages and blog entries, often represent the worst case. The data may need to be paraphrased, characterized, or summarized in a way that does not involve a direct transfer. As the use of business analytics grows (see Chapter 17), the use of automated ways to paraphrase or characterize these and other types of unstructured data grows, too.

Independent of the structure, collected data may exist in an electronic form that needs to be changed in order to be analyzed. For example, data presented as a digital picture of Excel worksheets would need to be changed into an actual Excel worksheet before that data could be analyzed. In this example, the electronic encoding of the data changes from a picture format such as jpeg to an Excel workbook format. Sometimes, individual numerical values that have been collected may need to be changed, especially collected values that result from a computational process. Demonstrate this issue in Excel by entering a formula that is equivalent to the expression $1 \times (0.5 - 0.4 - 0.1)$. This should evaluate as 0, but Excel evaluates to a very small negative number. Altering that value to 0 would be part of the data cleaning process.

Stacking and Unstacking Data

When collecting data for a numerical variable, subdividing that data into two or more groups for analysis may be necessary. For example, data about the cost of a restaurant meal in an urban area might be subdivided to consider the cost of meals at restaurants in the center city district separately from the meal costs at metro area restaurants. When using data that represent two or more groups, data can be arranged as either unstacked or stacked.

To use an **unstacked** arrangement, create separate numerical variables for each group. For this example, create a center city meal cost variable and a second variable to hold the meal costs at metro area restaurants. To use a **stacked** arrangement format, pair the single numerical variable meal cost with a second, categorical variable that contains two categories, such as center city and metro area. If collecting data for several numerical variables, each of which will be subdivided in the same way, stacking the data will be the more efficient choice.

When using software to analyze data, a specific procedure may require data to be stacked (or unstacked). When such cases arise using Microsoft Excel for problems or examples that this book discusses, a workbook or project will contain that data in both arrangements. For example, Restaurants, that Chapter 2 uses for several examples, contains both the original (stacked) data about restaurants as well as an unstacked worksheet (or data table) that contains the meal cost by location, center city or metro area.

Recoding Variables

After data have been collected, categories defined for a categorical variable may need to be reconsidered or a numerical variable may need to be transformed into a categorical variable by assigning individual numeric values to one of several groups. For either case, define a **recoded variable** that supplements or replaces the original variable in your analysis.

For example, having already defined the variable class standing with the categories freshman, sophomore, junior, and senior, a researcher decides to investigate the differences between lowerclassmen (freshmen or sophomores) and upperclassmen (juniors or seniors). The researcher can define a recoded variable UpperLower and assign the value Upper if a student is a junior or senior and assign the value Lower if the student is a freshman or sophomore.

When recoding variables, make sure that one and only one of the new categories can be assigned to any particular value being recoded and that each value can be recoded successfully by one of your new categories, the properties known as being **mutually exclusive** and **collectively exhaustive**.

When recoding numerical variables, pay particular attention to the operational definitions of the categories created for the recoded variable, especially if the categories are not self-defining ranges. For example, while the recoded categories Under 12, 12–20, 21–34, 35–54, and 55-and-over are self-defining for age, the categories child, youth, young adult, middle aged, and senior each need to be further defined in terms of mutually exclusive and collectively exhaustive numerical ranges.

PROBLEMS FOR SECTIONS 1.4 AND 1.5

APPLYING THE CONCEPTS

1.26 A study was conducted on the injuries sustained by workers in three different sections at a local factory. The following table shows the data for the first 5 cases out of a total of 25 cases.

Case No.	Section	Cause of Injury	Severity of Injury
1	A	Fall	3
	C	Auto	2
3	BB	Fall	6
4	B	Fall	q
5	V	Violence	9

a. Identify the type of irregularities in the data.
b. Clean the data and add the missing values.

1.27 The amount of monthly data usage by a sample of 10 cell phone users (in MB) was:

0.4, 2.7MB, 5.6, 4.3, 11.4, 26.8, 1.6, 1,079, 8.3, 4.2

Are there any potential irregularities in the data?

1.28 Consider the following information: Susan, 31 years old, weighs 81 kg; Connie, 27 years old, weighs 50 kg; and Alex, 63 years old, weighs 67 kg.
a. Use the unstacked format to organize the data.
b. Use the stacked format to organize the data.

1.29 A hotel management company runs 10 hotels in a resort area. The hotels have a mix of pricing—some hotels have budget-priced rooms, some have moderate-priced rooms, and some have deluxe-priced rooms. Data are collected that indicate the number of rooms that are occupied at each hotel on each day of a month. Explain how the 10 hotels can be recoded into these three price categories.

1.6 Types of Survey Errors

Collected data in the form of compiled responses from a survey must be verifed to ensure that the results can be used in a decision-making process. Verification begins by evaluating the validity of the survey to make sure the survey does not lack objectivity or credibility. To do this, evaluate the purpose of the survey, the reason the survey was conducted, and for whom the survey was conducted.

Having validated the objectivity and credibility of the survey, determine whether the survey was based on a probability sample (see Section 1.3). Surveys that use nonprobability samples are subject to serious biases that render their results useless for decision-making purposes. In the case of the Coca-Cola managers concerned about the "Pepsi Challenge" results (see page 47), the managers failed to reflect on the subjective nature of the challenge as well as the nonprobability sample that this survey used. Had the managers done so, they might not have been so quick to make the reformulation blunder that was reversed just weeks later.

Even after verification, surveys can suffer from any combination of the following types of survey errors: coverage error, nonresponse error, sampling error, or measurement error. Developers of well-designed surveys seek to reduce or minimize these types of errors, often at considerable cost.

Coverage Error

The key to proper sample selection is having an adequate frame. **Coverage error** occurs if certain groups of items are excluded from the frame so that they have no chance of being selected in the sample or if items are included from outside the frame. Coverage error results in a **selection bias**. If the frame is inadequate because certain groups of items in the population were not properly included, any probability sample selected will provide only an estimate of the characteristics of the frame, not the *actual* population.

Nonresponse Error

Not everyone is willing to respond to a survey. **Nonresponse error** arises from failure to collect data on all items in the sample and results in a **nonresponse bias**. Because a researcher cannot always assume that persons who do not respond to surveys are similar to those who do, researchers need to follow up on the nonresponses after a specified period of time. Researchers should make several attempts to convince such individuals to complete

the survey and possibly offer an incentive to participate. The follow-up responses are then compared to the initial responses in order to make valid inferences from the survey (see the Cochran, Groves et al., and Lohr sources). The mode of response the researcher uses, such as face-to-face interview, telephone interview, paper questionnaire, or computerized questionnaire, affects the rate of response. Personal interviews and telephone interviews usually produce a higher response rate than do mail surveys—but at a higher cost.

Sampling Error

When conducting a probability sample, chance dictates which individuals or items will or will not be included in the sample. **Sampling error** reflects the variation, or "chance differences," from sample to sample, based on the probability of particular individuals or items being selected in the particular samples.

When there is a news report about the results of surveys or polls in newspapers or on the Internet, there is often a statement regarding a margin of error, such as "the results of this poll are expected to be within ± 4 percentage points of the actual value." This **margin of error** is the sampling error. Using larger sample sizes reduces the sampling error. Of course, doing so increases the cost of conducting the survey.

Measurement Error

In the practice of good survey research, design surveys with the intention of gathering meaningful and accurate information. Unfortunately, the survey results are often only a proxy for the ones sought. Unlike height or weight, certain information about behaviors and psychological states is impossible or impractical to obtain directly.

When surveys rely on self-reported information, the mode of data collection, the respondent to the survey, or the survey itself can be possible sources of **measurement error**. Satisficing, social desirability, reading ability, and/or interviewer effects can be dependent on the mode of data collection. The social desirability bias or cognitive/memory limitations of a respondent can affect the results. Vague questions, double-barreled questions that ask about multiple issues but require a single response, or questions that ask the respondent to report something that occurs over time but fail to clearly define the extent of time about which the question asks (the reference period) are some of the survey flaws that can cause errors.

To minimize measurement error, standardize survey administration and respondent understanding of questions, but there are many barriers to this (Bremer, Fowler, Sudman).

Ethical Issues About Surveys

Ethical considerations arise with respect to the four types of survey error. Coverage error can result in selection bias and becomes an ethical issue if particular groups or individuals are purposely excluded from the frame so that the survey results are more favorable to the survey's sponsor. Nonresponse error can lead to nonresponse bias and becomes an ethical issue if the sponsor knowingly designs the survey so that particular groups or individuals are less likely than others to respond. Sampling error becomes an ethical issue if the findings are purposely presented without reference to sample size and margin of error so that the sponsor can promote a viewpoint that might otherwise be inappropriate. Measurement error can become an ethical issue in one of three ways: (1) a survey sponsor chooses leading questions that guide the respondent in a particular direction; (2) an interviewer, through mannerisms and tone, purposely makes a respondent obligated to please the interviewer or otherwise guides the respondent in a particular direction; or (3) a respondent willfully provides false information.

Ethical issues also arise when the results of nonprobability samples are used to form conclusions about the entire population. When using a nonprobability sampling method, explain the sampling procedures and state that the results cannot be generalized beyond the sample.

CONSIDER THIS

New Media Surveys/Old Survey Errors

Software company executives decide to create a "customer experience improvement program" to record how customers use the company's products, with the goal of using the collected data to make product enhancements. Product marketers decide to use social media websites to collect consumer feedback. These people risk making the same type of survey error that led to the quick demise of a very successful magazine nearly 80 years ago.

By 1935, "straw polls" conducted by the magazine *Literary Digest* had successfully predicted five consecutive U.S. presidential elections. For the 1936 election, the magazine promised its largest poll ever and sent about 10 million ballots to people all across the country. After tabulating more than 2.3 million ballots, the *Digest* confidently proclaimed that Alf Landon would be an easy winner over Franklin D. Roosevelt. The actual results: FDR won in a landslide, and Landon received the fewest electoral votes in U.S. history.

Being so wrong ruined the reputation of *Literary Digest*, and it would cease publication less than two years after it made its erroneous claim. A review much later found that the low response rate (less than 25% of the ballots distributed were returned) and nonresponse error (Roosevelt voters were less likely to mail in a ballot than Landon voters) were significant reasons for the failure of the *Literary Digest* poll (Squire).

The *Literary Digest* error proved to be a watershed event in the history of sample surveys. First, the error disproved the assertion that the larger the sample is, the better the results will be—an assertion some people still mistakenly make today. The error paved the way for the modern methods of sampling discussed in this chapter and gave prominence to the more "scientific" methods that George Gallup and Elmo Roper both used to correctly predict the 1936 elections. (Today's Gallup Polls and Roper Reports remember those researchers.)

In more recent times, Microsoft software executives overlooked that experienced users could easily opt out of participating in their improvement program. This created another case of nonresponse error that may have led to the improved product (Microsoft Office) being so poorly received initially by experienced Office users who, by being more likely to opt out of the improvement program, biased the data that Microsoft used to determine Office "improvements."

And while those product marketers may be able to collect a lot of customer feedback data, those data also suffer from nonresponse error. In collecting data from social media websites, the marketers cannot know who chose *not* to leave comments. The marketers also cannot verify if the data collected suffer from a selection bias due to a coverage error.

That you might use media newer than the mailed, dead-tree form that *Literary Digest* used does not mean that you automatically avoid the old survey errors. Just the opposite—the accessibility and reach of new media makes it much easier for unknowing people to commit such errors.

PROBLEMS FOR SECTION 1.6

APPLYING THE CONCEPTS

1.30 A survey indicates that the vast majority of college students own their own smartphones. What information would you want to know before you accepted the results of this survey?

1.31 An online survey was conducted to determine the overall perception of a workshop conducted for a group of 150 employees at a corporate company. Each employee has to complete the survey in 10 minutes. However, only those with Internet access will be able to complete the survey. Of these employees, only 80% took the survey and some completed it in less than 2 minutes. Identify a possible example of the following errors.

a. Coverage error.
b. Sampling error.
c. Nonresponse error.
d. Measurement error.

SELF TEST **1.32** The librarian at a university library in Norway conducts a survey to measure the students' satisfaction regarding the library's services. Across a period of 3 weeks, he interviews every 50th student who enters the library. He explains the various services the library offers before recording the participants' responses. Identify *potential* ethical concerns or errors in the survey.

1.33 A 2019 PwC survey of 1,000 U.S. executives (available at **pwc.to/2DAHI4q**) indicated that artificial intelligence (AI) is no longer seen as a side project or science experiment. Eighty percent of U.S. CEOs think AI will significantly change the way they will do business in the next five years. At the same time, they are concerned about AI risks that could undermine investments. What risks concern CEOs most? Forty-three percent cite new privacy threats. But CEOs also note growing concerns over how AI could affect cybersecurity,

employment, inequality, and the environment. A majority of CEOs are already taking steps to address these concerns by developing and deploying AI systems that are trustworthy.

What additional information would you want to know about the survey before you accepted the results for the study?

1.34 A recent survey (avialable at **bit.ly/2H5mQ6g**) points to the transformation underway in the automotive retail landscape. The 2019 KPMG Global Automotive Executive Study found that automobile executives believe the number of physical automotive retail outlets, as we know them today, will be reduced by 30% to 50%. Eighty-two percent of automobile executives strongly agree that the only viable option for physical retail outlets will be the transformation into service factories, used car hubs, or focusing on an ID-management approach, where the customer is recognized at every single touchpoint.

What additional information would you want to know about the survey before you accepted the results of the study?

USING **STATISTICS**
Defining Moments, Revisited

The New Coke and airline quality cases illustrate missteps that can occur during the define and collect tasks of the DCOVA framework. To use statistics effectively, you must properly define a business problem or goal and then collect data that will allow you to make observations and reach conclusions that are relevant to that problem or goal.

In the New Coke case, managers failed to consider that data collected about a taste test would not necessarily provide useful information about the sales issues they faced. The managers also did not realize that the test used improper sampling techniques, deliberately introduced biases, and were subject to coverage and nonresponse errors. Those mistakes invalidated the test, making the conclusion that New Coke tasted better than Pepsi an invalid claim.

In the airline quality case, no mistakes in defining and collecting data were made. The results that fliers like quality was a valid one, but decision makers overlooked that quality was not the most significant factor for people buying seats on transatlantic flights (price was). This case illustrates that no matter how well you apply statistics, if you do not properly analyze the business problem or goal being considered, you may end up with valid results that lead you to invalid management decisions.

SUMMARY

In this chapter, you learned the details about the Define and Collect tasks of the DCOVA framework, which are important first steps to applying statistics properly to decision making. You learned that defining variables means developing an operational definition that includes establishing the type of variable and the measurement scale that the variable uses. You learned important details about data collection as well as some new basic vocabulary terms (sample, population, and parameter) and a more precise definition of statistic. You specifically learned about sampling and the types of sampling methods available to you. Finally, you surveyed data preparation considerations and learned about the type of survey errors you can encounter.

REFERENCES

Biemer, P. B., R. M. Graves, L. E. Lyberg, A. Mathiowetz, and S. Sudman. *Measurement Errors in Surveys*. New York: Wiley Interscience, 2004.

Cochran, W. G. *Sampling Techniques*, 3rd ed. New York: Wiley, 1977.

Fowler, F. J. *Improving Survey Questions: Design and Evaluation*, *Applied Special Research Methods Series*, Vol. 38, Thousand Oaks, CA: Sage Publications, 1995.

Groves R. M., F. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau. *Survey Methodology*, 2nd ed. New York: John Wiley, 2009.

Hellerstein, J. "Quantitative Data Cleaning for Large Databases." **bit.ly/2q7PGIn**.

Lohr, S. L. *Sampling Design and Analysis*, 2nd ed. Boston, MA: Brooks/Cole Cengage Learning, 2010.

Polaris Marketing Research. "Brilliant Marketing Research or What? The New Coke Story," posted September 20, 2011. **bit.ly/1DofHSM** (removed).

Press, G. "Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says," posted March 23, 2016. **bit.ly/2oNCwzh**.

Rosenbaum, D. "The New Big Data Magic," posted August 20, 2011. **bit.ly/1DUMWzv**.

Osbourne, J. *Best Practices in Data Cleaning*. Thousand Oaks, CA: Sage Publications, 2012.

Squire, P. "Why the 1936 *Literary Digest* Poll Failed." *Public Opinion Quarterly* 52 (1988): 125–133.

Sudman, S., N. M. Bradburn, and N. Schwarz. *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco, CA: Jossey-Bass, 1993.

KEY TERMS

categorical variable 47
cluster 52
cluster sample 52
collectively exhaustive 56
continuous variable 48
convenience sample 51
coverage error 57
data cleaning 54
discrete variable 48
frame 51
interval scale 48
judgment sample 51
margin of error 58
measurement error 58
measurement scale 48
missing value 55
mutually exclusive 56
nominal scale 48
nonprobability sample 51
nonresponse bias 57
nonresponse error 57
numerical variable 47
operational definition 47
ordinal scale 48
outlier 55
parameter 50
population 49
primary data source 50
probability sample 51
ratio scale 48
recoded variable 56
sample 49
sampling error 58
sampling with replacement 51
sampling without replacement 51
secondary data source 50
selection bias 57
simple random sample 51
stacked 56
statistic 50
strata 52
stratified sample 52
systematic sample 52
table of random numbers 52
treatment 50
unstacked 56

CHECKING YOUR UNDERSTANDING

1.35 What is the difference between a sample and a population?

1.36 What is the difference between a statistic and a parameter?

1.37 What is the difference between a categorical variable and a numerical variable?

1.38 What is the difference between a discrete numerical variable and a continuous numerical variable?

1.39 State one example of a nominal and an ordinal scaled variable that might be represented in the same numerical form.

1.40 What is the difference between an interval scaled variable and a ratio scaled variable?

1.41 State one advantage of each of the probability sampling methods.

1.42 What is the difference between a missing value and an outlier?

1.43 What is the difference between unstacked and stacked variables?

1.44 What is the difference between coverage error and nonresponse error?

1.45 What is the difference between sampling error and measurement error?

CHAPTER REVIEW PROBLEMS

1.46 Visit the official Microsoft Excel product website, **products.office.com/excel**. Review the features of the program you chose and then state the ways the program could be useful in statistical analysis.

1.47 Results of a 2017 Computer Services, Inc. (CSI) survey of a sample of 163 bank executives reveal insights on banking priorities among financial institutions (**goo.gl/mniYMM**). As financial institutions begin planning for a new year, of utmost importance is boosting profitability and identifying growth areas. The results show that 55% of bank institutions note customer experience initiatives as an area in which spending is expected to increase. Implementing a customer relationship management (CRM) solution was ranked as the top most important omnichannel strategy to pursue with 41% of institutions citing digital banking enhancements as the greatest anticipated strategy to enhance the customer experience.

a. Describe the population of interest.
b. Describe the sample that was collected.
c. Describe a parameter of interest.
d. Describe the statistic used to estimate the parameter in (c).

1.48 The Gallup organization releases the results of recent polls on its website, **www.gallup.com**. Visit this site and read an article of interest.
a. Describe the population of interest.
b. Describe the sample that was collected.
c. Describe a parameter of interest.
d. Describe the statistic used to estimate the parameter in (c).

1.49 A 2019 PwC survey of 1,000 U.S. executives indicated that artificial intelligence (AI) is no longer seen as a side project. Eighty percent of U.S. CEOs think AI will significantly change the way they will do business in the next five years. At the same time, these CEOs are concerned about AI risks that could undermine investments. What risks concern CEOs most? Forty-three percent cite new privacy threats. But CEOs also note growing concerns over how AI could affect cybersecurity, employment, inequality, and the environment. A majority of CEOs are already taking steps to address these concerns by developing and deploying AI systems that are trustworthy.

Source: "US CEO agenda 2019," PwC, **pwc.to/2UuoVAX**.
a. Describe the population of interest.
b. Describe the sample that was collected.
c. Describe a parameter of interest.
d. Describe the statistic used to estimate the parameter in (c).

1.50 The American Community Survey (**www.census.gov/acs**) provides data every year about communities in the United States. Addresses are randomly selected and respondents are required to supply answers to a series of questions.
a. Describe a variable for which data are collected.
b. Is the variable categorical or numerical?
c. If the variable is numerical, is it discrete or continuous?

1.51 Examine Zarca Interactive's "Sample Employee Satisfaction Survey/Sample Questions for Employee Satisfaction Survey," available at **bit.ly/21qjI6F**.
a. Give an example of a categorical variable included in the survey.
b. Give an example of a numerical variable included in the survey.

1.52 Three professors examined awareness of four widely disseminated retirement rules among employees at the University of Utah. These rules provide simple answers to questions about retirement planning (R. N. Mayer, C. D. Zick, and M. Glaittle, "Public Awareness of Retirement Planning Rules of Thumb," *Journal of Personal Finance*, 2011 10(62), 12–35). At the time of the investigation, there were approximately 10,000 benefited employees, and 3,095 participated in the study. Demographic data collected on these 3,095 employees included gender, age (years), education level (years completed), marital status, household income ($), and employment category.
a. Describe the population of interest.
b. Describe the sample that was collected.
c. Indicate whether each of the demographic variables mentioned is categorical or numerical.

1.53 Social media provides an enormous amount of data about the activities and habits of people using social platforms like Facebook and Twitter. The belief is that mining that data provides a treasure trove for those who seek to quantify and predict future human behavior. A marketer is planning a survey of Internet users in the United States to determine social media usage. The objective of the survey is to gain insight on these three items: key social media platforms used, frequency of social media usage, and demographics of key social media platform users.
a. For each of the three items listed, indicate whether the variables are categorical or numerical. If a variable is numerical, is it discrete or continuous?
b. Develop five categorical questions for the survey.
c. Develop five numerical questions for the survey.

CHAPTER 1

CASES

Managing Ashland MultiComm Services

Ashland MultiComm Services (AMS) provides high-quality telecommunications services in the Greater Ashland area. AMS traces its roots to a small company that redistributed the broadcast television signals from nearby major metropolitan areas but has evolved into a provider of a wide range of broadband services for residential customers.

AMS offers subscription-based services for digital cable television, local and long-distance telephone services, and high-speed Internet access. Recently, AMS has faced competition from other service providers as well as Internet-based, on-demand streaming services that have caused many customers to "cut the cable" and drop their subscription to cable video services.

AMS management believes that a combination of increased promotional expenditures, adjustment in subscription fees, and improved customer service will allow AMS to successfully face these challenges. To help determine the proper mix of strategies to be taken, AMS management has decided to organize a research team to undertake a study.

The managers suggest that the research team examine the company's own historical data for number of subscribers, revenues, and subscription renewal rates for the past few years. They direct the team to examine year-to-date data as well because the managers suspect that some of the changes they have seen have been a relatively recent phenomena.

1. What type of data source would the company's own historical data be? Identify other possible data sources that the research

team might use to examine the current marketplace for residential broadband services in a city such as Ashland.

2. What type of data collection techniques might the team employ?
3. In their suggestions and directions, the AMS managers have named a number of possible variables to study but offered no operational definitions for those variables. What types of possible misunderstandings could arise if the team and managers do not first properly define each variable cited?

CardioGood Fitness

CardioGood Fitness is a developer of high-quality cardiovascular exercise equipment. Its products include treadmills, fitness bikes, elliptical machines, and e-glides. CardioGood Fitness looks to increase the sales of its treadmill products and has hired The AdRight Agency, a small advertising firm, to create and implement an advertising program. The AdRight Agency plans to identify particular market segments that are most likely to buy their clients' goods and services and then locate advertising outlets that will reach that market group. This activity includes collecting data on clients' actual sales and on the customers who make the purchases, with the goal of determining whether there is a distinct profile of the typical customer for a particular product or service. If a distinct profile emerges, efforts are made to match that profile to advertising outlets known to reflect the particular profile, thus targeting advertising directly to high-potential customers.

CardioGood Fitness sells three different lines of treadmills. The TM195 is an entry-level treadmill. It is as dependable as other models offered by CardioGood Fitness, but with fewer programs and features. It is suitable for individuals who thrive on minimal programming and the desire for simplicity to initiate their walk or hike. The TM195 sells for $1,500.

The middle-line TM498 adds to the features of the entry-level model: two user programs and up to 15% elevation upgrade. The TM498 is suitable for individuals who are walkers at a transitional stage from walking to running or midlevel runners. The TM498 sells for $1,750.

The top-of-the-line TM798 is structurally larger and heavier and has more features than the other models. Its unique features include a bright blue backlit LCD console, quick speed and incline keys, a wireless heart rate monitor with a telemetric chest strap, remote speed and incline controls, and an anatomical figure that specifies which muscles are minimally and maximally activated. This model features a nonfolding platform base that is designed to handle rigorous, frequent running; the TM798 is therefore appealing to someone who is a power walker or a runner. The selling price is $2,500.

As a first step, the market research team at AdRight is assigned the task of identifying the profile of the typical customer for each treadmill product offered by CardioGood Fitness. The market research team decides to investigate whether there are differences across the product lines with respect to customer characteristics. The team decides to collect data on individuals who purchased a treadmill at a CardioGood Fitness retail store during the prior three months.

The team decides to use both business transactional data and the results of a personal profile survey that every purchaser completes as the team's sources of data. The team identifies the following customer variables to study: product purchased—TM195, TM498, or TM798; gender; age, in years; education, in years; relationship status, single or partnered; annual household income ($); mean number of times the customer plans to use the treadmill each week; mean number of miles the customer expects to walk/run each week; and self-rated fitness on a 1-to-5 scale, where 1 is poor shape and 5 is excellent shape. For this set of variables:

1. Which variables in the survey are categorical?
2. Which variables in the survey are numerical?
3. Which variables are discrete numerical variables?

Clear Mountain State Student Survey

The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students who attend CMSU. They create and distribute a survey of 14 questions and receive responses from 111 undergraduates (stored in Student Survey).

Download (see Appendix C) and review the survey document **CMUndergradSurvey.pdf**. For each question asked in the survey, determine whether the variable is categorical or numerical. If you determine that the variable is numerical, identify whether it is discrete or continuous.

Learning With the Digital Cases

Identifying and preventing misuses of statistics is an important responsibility for all managers. The Digital Cases allow you to practice the skills necessary for this important task.

Each chapter's Digital Case tests your understanding of how to apply an important statistical concept taught in the chapter. As in many business situations, not all of the information you encounter will be relevant to your task, and you may occasionally discover conflicting information that you have to resolve in order to complete the case.

To assist your learning, each Digital Case begins with a learning objective and a summary of the problem or issue at hand. Each case directs you to the information necessary to reach your own conclusions and to answer the case questions. Many cases, such as the sample case worked out next, extend a chapter's Using Statistics scenario. You can download digital case files, which are PDF format documents that may contain extended features such as interactivity or data file attachments. Open these files with a current version of Adobe Reader, as other PDF programs may not support the extended features. (For more information, see Appendix C.)

To illustrate learning with a Digital Case, open the Digital Case file **WhitneyWireless.pdf** that contains summary information about the Whitney Wireless business. Apparently, from the claim on the title page, this business is celebrating its "best sales year ever."

Review the **Who We Are**, **What We Do**, and **What We Plan to Do** sections on the second page. Do these sections contain any useful information? What *questions* does this passage raise? Did you notice that while many facts are presented, no data that would support the claim of "best sales year ever" are presented? And were those mobile "mobilemobiles" used solely

for promotion? Or did they generate any sales? Do you think that a talk-with-your-mouth-full event, however novel, would be a success?

Continue to the third page and the **Our Best Sales Year Ever!** section. How would you support such a claim? With a table of numbers? Remarks attributed to a knowledgeable source? Whitney Wireless has used a chart to present "two years ago" and "latest twelve months" sales data by category. Are there any problems with what the company has done? *Absolutely!*

Take a moment to identify and reflect on those problems. Then turn to pages 4 though 6 that present an annotated version of the first three pages and discusses some of the problems with this document.

In subsequent Digital Cases, you will be asked to provide this type of analysis, using the open-ended case questions as your guide. Not all the cases are as straightforward as this example, and some cases include perfectly appropriate applications of statistical methods. And none have annotated answers!