Pearson New International Edition

Statistical Methods for the Social Sciences

Alan Agresti Barbara Finlay

Fourth Edition

ALWAYS LEARNING

PEARSON

CHAPTER 1

Introduction

1.1 INTRODUCTION TO STATISTICAL METHODOLOGY

The past quarter-century has seen a dramatic increase in the use of statistical methods in the social sciences. There are many reasons for this. More research in the social sciences has taken on a quantitative orientation. Like research in other sciences, research in the social sciences often studies questions of interest by analyzing evidence provided by empirical data. The growth of the Internet has resulted in an increase in the amount of readily available quantitative information. Finally, with the evolution of evermore powerful computers, software, and statistical methodology, new methods are available that can more realistically address the questions that arise in social science research.

Why Study Statistics?

The increased use of statistics is evident in the changes in the content of articles published in social science research journals and reports prepared in government and private industry. A quick glance through recent issues of journals such as *American Political Science Review* and *American Sociological Review* reveals the fundamental role of statistics in research. For example, to learn about which factors have the greatest impact on student performance in school or to investigate which factors affect people's political beliefs or the quality of their health care or their decision about when to retire, researchers collect information and process it using statistical analyses. Because of the role of statistics in many research studies, more and more academic departments require that their majors take statistics courses.

These days, social scientists work in a wide variety of areas that use statistical methods, such as governmental agencies, business organizations, and health care facilities. For example, social scientists in government agencies dealing with human welfare or environmental issues or public health policy invariably need to use statistical methods or at least read reports that contain statistics. Medical sociologists often must evaluate recommendations from studies that contain quantitative investigations of new therapies or new ways of caring for the elderly. Some social scientists help managers to evaluate employee performance using quantitative benchmarks and to determine factors that help predict sales of products. In fact, increasingly many jobs for social scientists expect a knowledge of statistical methods as a basic work tool. As the joke goes, "What did the sociologist who passed statistics say to the sociologist who failed it? 'I'll have a Big Mac, fries, and a Coke.' "

But an understanding of statistics is important even if you never use statistical methods in your career. Every day you are exposed to an explosion of information, from advertising, news reporting, political campaigning, surveys about opinions on controversial issues, and other communications containing statistical arguments. Statistics helps you make sense of this information and better understand the world. You will find concepts from this text helpful in judging the information you will encounter in your everyday life.

We realize you are not reading this book in hopes of becoming a statistician. In addition, you may suffer from math phobia and feel fear at what lies ahead. Please be assured that you can read this book and learn the primary concepts and methods of statistics with little knowledge of mathematics. Just because you may have had difficulty in math courses before does not mean you will be at a disadvantage here. To understand this book, logical thinking and perseverance are more important than mathematics. In our experience, the most important factor in how well you do in a statistics course is how much time you spend on the course—attending class, doing homework, reading and re-reading this text, studying your class notes, working together with your fellow students, getting help from your professor or teaching assistant—not your mathematical knowledge or your gender or your race or whether you feel fear of statistics at the beginning.

Don't be frustrated if learning comes slowly and you need to read a chapter a few times before it starts to make sense. Just as you would not expect to take a single course in a foreign language and be able to speak that language fluently, the same is true with the language of statistics. Once you have completed even a portion of this text, however, you will better understand how to make sense of statistical information.

Data

Information gathering is at the heart of all sciences, providing the ***observations*** used in statistical analyses. The observations gathered on the characteristics of interest are collectively called ***data***.

For example, a study might conduct a survey of 1000 people to observe characteristics such as opinion about the legalization of marijuana, political party affiliation, political ideology, how often attend religious services, number of years of education, annual income, marital status, race, and gender. The data for a particular person would consist of observations such as (opinion = do not favor legalization, party = Republican, ideology = conservative, religiosity = once a week, education = 14 years, annual income in range 40–60 thousand dollars, marital status = married, race = white, gender = female). Looking at the data in the right way helps us learn about how such characteristics are related. We can then answer questions such as, "Do people who attend church more often tend to be more politically conservative?"

To generate data, the social sciences use a wide variety of methods, including surveys, experiments, and direct observation of behavior in natural settings. In addition, social scientists often analyze data already recorded for other purposes, such as police records, census materials, and hospital files. Existing archived collections of data are called ***databases***. Many databases are now available on the Internet. A very important database for social scientists contains results since 1972 of the General Social Survey.

EXAMPLE 1.1 The General Social Survey (GSS)

Every other year, the National Opinion Research Center at the University of Chicago conducts the General Social Survey (GSS). This survey of about 2000 adults provides data about opinions and behaviors of the American public. Social scientists use it to investigate how adult Americans answer a wide diversity of questions, such as, "Do you believe in life after death?," "Would you be willing to pay higher prices in order

to protect the environment?," and "Do you think a preschool child is likely to suffer if his or her mother works?" Similar surveys occur in other countries, such as the General Social Survey administered by Statistics Canada, the British Social Attitudes Survey, and the Eurobarometer survey and European Social Survey for nations in the European Union.

It is easy to get summaries of data from the GSS database. We'll demonstrate, using a question it asked in one survey, "About how many good friends do you have?"

- Go to the Web site sda.berkeley.edu/GSS/ at the Survey Documentation and Analysis site at the University of California, Berkeley.
- Click on *New SDA*.
- The GSS name for the question about number of good friends is NUMFREND. Type NUMFREND as the *Row* variable name. Click on *Run the table*.

Now you'll see a table that shows the possible values for 'number of good friends' and the number of people and the percentage who made each possible response. The most common responses were 2 and 3 (about 16% made each of these responses). ■

What Is Statistics?

In this text, we use the term "statistics" in the broad sense to refer to methods for obtaining and analyzing data.

Statistics

Statistics consists of a body of methods for obtaining and analyzing data.

Specifically, statistics provides methods for

1. ***Design***: Planning how to gather data for research studies
2. ***Description***: Summarizing the data
3. ***Inference***: Making predictions based on the data

Design refers to planning how to obtain the data. For a survey, for example, the design aspects would specify how to select the people to interview and would construct the questionnaire to administer.

Description refers to summarizing data, to help understand the information they provide. For example, an analysis of the number of good friends based on the GSS data might start with a list of the number reported for each of the people who responded to that question that year. The raw data are a complete listing of observations, person by person. These are not easy to comprehend, however. We get bogged down in numbers. For presentation of results, instead of listing *all* observations, we could summarize the data with a graph or table showing the percentages reporting 1 good friend, 2 good friends, 3, . . ., and so on. Or we could report the average number of good friends, which was 6, or the most common response, which was 2. Graphs, tables and numerical summaries are called ***descriptive statistics***.

Inference refers to making predictions based on data. For instance, for the GSS data on reported number of good friends, 6.2% reported having only 1 good friend. Can we use this information to predict the percentage of the more than 200 million adults in the U.S. at that time who had only 1 good friend? A method presented in this book allows us to predict that that percentage is no greater than 8%. Predictions made using data are called ***statistical inferences***.

Description and ***inference*** are the two types of ***statistical analysis***—ways of analyzing the data. Social scientists use descriptive and inferential statistics to answer questions about social phenomena. For instance, "Is having the death penalty

available for punishment associated with a reduction in violent crime?" "Does student performance in schools depend on the amount of money spent per student, the size of the classes, or the teachers' salaries?"

1.2 DESCRIPTIVE STATISTICS AND INFERENTIAL STATISTICS

Section 1.1 explained that statistics consists of methods for *designing* studies and *analyzing* data collected in the studies. Methods for analyzing data include descriptive methods for summarizing the data and inferential methods for making predictions. A statistical analysis is classified as ***descriptive*** or ***inferential***, according to whether its main purpose is to describe the data or to make predictions. To explain this distinction further, we next define the *population* and the *sample*.

Populations and Samples

The entities that a study observes are called the ***subjects*** for the study. Usually the subjects are people, such as in the GSS, but they might instead be families, schools, cities, or companies, for instance.

Population and Sample

The ***population*** is the total set of subjects of interest in a study. A ***sample*** is the subset of the population on which the study collects data.

In the 2004 GSS, the sample was the 2813 adult Americans who participated in the survey. The population was all adult Americans at that time—more than 200 million people.

The ultimate goal of any study is to learn about populations. But it is almost always necessary, and more practical, to observe only samples from those populations. For example, the GSS and polling organizations such as the Gallup poll usually select samples of about 1000–3000 Americans to collect information about opinions and beliefs of the population of *all* Americans.

Descriptive Statistics

Descriptive statistics summarize the information in a collection of data.

Descriptive statistics consist of graphs, tables, and numbers such as averages and percentages. The main purpose of descriptive statistics is to reduce the data to simpler and more understandable forms without distorting or losing much information.

Although data are usually available only for a sample, descriptive statistics are also useful when data are available for the entire population, such as in a census. By contrast, inferential statistics apply when data are available only for a sample but we want to make a prediction about the entire population.

Inferential Statistics

Inferential statistics provide predictions about a population, based on data from a sample of that population.

EXAMPLE 1.2 Belief in Heaven

In two of its surveys, the GSS asked, "Do you believe in heaven?" The population of interest was the collection of all adults in the United States. In the most recent survey

in which this was asked, 86% of the 1158 sampled subjects answered *yes*. We would be interested, however, not only in those 1158 people but in the *entire population* of all adults in the U.S.

Inferential statistics provide a prediction about the larger population using the sample data. An inferential method presented in Chapter 5 predicts that the population percentage that believe in heaven falls between 84% and 88%. That is, the sample value of 86% has a "margin of error" of 2%. Even though the sample size was tiny compared to the population size, we can conclude that a large percentage of the population believed in heaven. ■

Inferential statistical analyses can predict characteristics of entire populations quite well by selecting samples that are small relative to the population size. That's why many polls sample only about a thousand people, even if the population has millions of people. In this book, we'll see why this works.

In the past quarter-century, social scientists have increasingly recognized the power of inferential statistical methods. Presentation of these methods occupies a large portion of this textbook, beginning in Chapter 5.

Parameters and Statistics

Parameters and Statistics

A ***parameter*** is a numerical summary of the population. A ***statistic*** is a numerical summary of the sample data.

Example 1.2 estimated the percentage of Americans who believe in heaven. The parameter was the population percentage who believed in heaven. Its value was unknown. The inference about this parameter was based on a statistic—the percentage of the 1158 subjects interviewed in the survey who answered *yes*, namely, 86%. Since this number *describes* a characteristic of the sample, it is a descriptive statistic.

In practice, the main interest is in the values of the parameters, not the values of the statistics for the particular sample selected. For example, in viewing results of a poll before an election, we're more interested in the *population* percentages favoring the various candidates than in the *sample* percentages for the people interviewed. The sample and statistics describing it are important only insofar as they help us make inferences about unknown population parameters.

An important aspect of statistical inference involves reporting the likely *precision* of the sample statistic that estimates the population parameter. For Example 1.2 on belief in heaven, an inferential statistical method predicted how close the *sample* value of 86% was likely to be to the unknown percentage of the *population* believing in heaven. The reported margin of error was 2%.

When data exist for an entire population, such as in a census, it's possible to find the actual values of the parameters of interest. Then there is no need to use inferential statistical methods.

Defining Populations: Actual and Conceptual

Usually the population to which inferences apply is an actual set of subjects. In Example 1.2, it was adult residents of the U.S. Sometimes, though, the generalizations refer to a *conceptual* population—one that does not actually exist but is hypothetical.

For example, suppose a consumer organization evaluates gas mileage for a new model of an automobile by observing the average number of miles per gallon for five sample autos driven on a standardized 100-mile course. Their inferences refer to the performance on this course for the conceptual population of *all* autos of this model that will be or could hypothetically be manufactured.

1.3 THE ROLE OF COMPUTERS IN STATISTICS

Over time, ever more powerful computers reach the market, and powerful and easy-to-use software is further developed for statistical methods. This software provides an enormous boon to the use of statistics.

Statistical Software

SPSS (Statistical Package for the Social Sciences), SAS, MINITAB, and Stata are the most popular statistical software on college campuses. It is much easier to apply statistical methods using these software than using hand calculation. Moreover, many methods presented in this text are too complex to do by hand or with hand calculators.

Most chapters of this text, including all those that present methods requiring considerable computation, show examples of the output of statistical software. One purpose of this textbook is to teach you what to look for in output and how to interpret it. Knowledge of computer programming is not necessary for using statistical software or for reading this book.

The text appendix explains how to use SPSS and SAS, organized by chapter. You can refer to this appendix as you read each chapter to learn how to use them to perform the analyses of that chapter.

Data Files

Figure 1.1 shows an example of data organized in a ***data file*** for analysis by statistical software. A data file has the form of a spreadsheet:

- Any one row contains the observations for a particular subject in the sample.
- Any one column contains the observations for a particular characteristic.

Figure 1.1 is a window for editing data in SPSS. It shows data for the first ten subjects in a sample, for the characteristics sex, racial group, marital status, age, and annual income (in thousands of dollars). Some of the data are numerical, and some consist of labels. Chapter 2 introduces the types of data for data files.

Uses and Misuses of Statistical Software

A note of caution: The easy access to statistical methods using software has dangers as well as benefits. It is simple to apply inappropriate methods. A computer performs the analysis requested whether or not the assumptions required for its proper use are satisfied.

Incorrect analyses result when researchers take insufficient time to understand the statistical method, the assumptions for its use, or its appropriateness for the specific problem. It is vital to understand the method before using it. Just knowing how to use statistical software does not guarantee a proper analysis. You'll need a good background in statistics to understand which method to select, which options to choose in that method, and how to make valid conclusions from the output. The main purpose of this text is to give you this background.

chapter_1_data.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

1 : sex | female

	subject	sex	race	married	age	income
1	1	female	white	yes	23	18.3
2	2	female	black	no	37	31.9
3	3	male	white	yes	47	64.0
4	4	female	white	yes	61	46.2
5	5	male	hispanic	yes	30	16.5
6	6	male	white	no	21	14.0
7	7	male	white	yes	55	26.1
8	8	female	white	no	27	59.8
9	9	female	hispanic	yes	61	21.5
10	10	male	black	no	47	50.0

Data View | Variable View

SPSS Processor is ready

FIGURE 1.1: Example of Part of a SPSS Data File

1.4 CHAPTER SUMMARY

The field of statistics includes methods for

- designing research studies,
- describing the data, and
- making inferences (predictions) using the data.

Statistical methods normally are applied to observations in a ***sample*** taken from the ***population*** of interest. ***Statistics*** summarize sample data, while ***parameters*** summarize entire populations. There are two types of statistical analyses:

- ***Descriptive statistics*** summarize sample or population data with numbers, tables, and graphs.
- ***Inferential statistics*** make predictions about population parameters, based on sample data.

A ***data file*** has a separate row of data for each subject and a separate column for each characteristic. Statistical methods are easy to apply to data files using software. This relieves us of computational drudgery and helps us focus on the proper application and interpretation of the methods.

PROBLEMS

Practicing the Basics

1.1. The Environmental Protection Agency (EPA) uses a few new automobiles of each brand every year to collect data on pollution emission and gasoline mileage performance. For the Toyota Prius brand, identify the (a) subject, (b) sample, (c) population.

1.2. In the 2006 gubernatorial election in California, an exit poll sampled 2705 of the 7 million people who voted. The poll stated that 56.5%

reported voting for the Republican candidate, Arnold Schwarzenegger. Of all 7 million voters, 55.9% voted for Schwarzenegger.

(a) For this exit poll, what was the population and what was the sample?

(b) Identify a statistic and a parameter.

1.3. The student government at the University of Wisconsin conducts a study about alcohol abuse among students. One hundred of the 40,858 members of the student body are sampled and asked to complete a questionnaire. One question asked is, "On how many days in the past week did you consume at least one alcoholic drink?"

(a) Identify the population of interest.

(b) For the 40,858 students, one characteristic of interest was the percentage who would respond *zero* to this question. This value is computed for the 100 students sampled. Is it a parameter or a statistic? Why?

1.4. The Institute for Public Opinion Research at Florida International University has conducted the FIU/Florida Poll (www.fiu.edu/orgs/ipor/ffp) of about 1200 Floridians annually since 1988 to track opinions on a wide variety of issues. The poll reported in 2006 that 67% of Floridians believe that state government should not make laws restricting access to abortion. Is 67% the value of a statistic, or of a parameter? Why?

1.5. A GSS asked subjects whether astrology—the study of star signs—has some scientific truth (GSS question SCITEST3). Of 1245 sampled subjects, 651 responded *definitely or probably true*, and 594 responded *definitely or probably not true*. The proportion responding *definitely or probably true* was $651/1245 = 0.523$.

(a) Describe the population of interest.

(b) For what population parameter might we want to make an inference?

(c) What sample statistic could be used in making this inference?

(d) Does the value of the statistic in (c) necessarily equal the parameter in (b)? Explain.

1.6. Go to the GSS Web site, sda.berkeley.edu/GSS/. By entering TVHOURS as the *Row variable*, find a summary of responses to the question, "On a typical day, about how many hours do you personally watch television?"

(a) What was the most common response?

(b) Is your answer in (a) a descriptive statistic, or an inferential statistic?

1.7. Go to the GSS Web site, sda.berkeley.edu/GSS/. By entering HEAVEN as the *Row variable*, you can find the percentages of people who said *definitely yes*, *probably yes*, *probably not*, and *definitely not* when asked whether they believed in heaven.

(a) Report the percentage who gave one of the *yes* responses.

(b) To obtain data for a particular year such as 1998, enter YEAR(1998) in the *Selection filter* option box before you click on *Run the Table*. Do this for HEAVEN in 1998, and report the percentage who gave one of the *yes* responses. (This question was asked only in 1991 and 1998.)

(c) Summarize opinions in 1998 about belief in hell (variable HELL in the GSS). Was the percentage of *yes* responses higher for HEAVEN or HELL?

1.8. The Current Population Survey (CPS) is a monthly survey of households conducted by the U.S. Census Bureau. A CPS of 60,000 households indicated that of those households, 8.1% of the whites, 22.3% of the blacks, 20.9% of the Hispanics, and 10.2% of the Asians had annual income below the poverty level (*Statistical Abstract of the United States, 2006*).

(a) Are these numbers statistics, or parameters? Explain.

(b) A method from this text predicts that the percentage of *all* black households in the United States having income below the poverty level is at least 21% but no greater than 24%. What type of statistical method does this illustrate—descriptive or inferential? Why?

1.9. A BBC story (September 9, 2004) about a poll in 35 countries concerning whether people favored George W. Bush or John Kerry in the 2004 U.S. Presidential election stated that Kerry was clearly preferred. Of the sample from Germany, 74% preferred Kerry, 10% preferred Bush, with the rest undecided or not responding. Multiple choice: The results for Germany are an example of

(a) descriptive statistics for a sample.

(b) inferential statistics about a population.

(c) a data file.

(d) a population.

1.10. Construct a data file describing the criminal behavior of five inmates in a local prison. The characteristics measured were race (with observations for the five subjects: white, black, white, Hispanic, white), age (19, 23, 38, 20, 41), length of sentence in years (2, 1, 10, 2, 5), whether convicted on a felony (no, no, yes, no, yes), number of prior arrests (values 2, 0, 8, 1, 5), number of prior convictions (1, 0, 3, 1, 4).

Concepts and Applications

1.11. The "Student survey" data file at *www.stat.ufl.edu/~aa/social/data.html*

shows responses of a class of social science graduate students at the University of Florida to a questionnaire that asked about GE = gender, AG = age in years, HI = high school GPA (on a four-point scale), CO = college GPA, DH = distance (in miles) of the campus from your home town, DR = distance (in miles) of the classroom from your current residence, NE = number of times a week you read a newspaper, TV = average number of hours per week that you watch TV, SP = average number of hours per week that you participate in sports or have other physical exercise, VE = whether you are a vegetarian (yes, no), AB = opinion about whether abortion should be legal in the first three months of pregnancy (yes, no), PI = political ideology (1 = very liberal, 2 = liberal, 3 = slightly liberal, 4 = moderate, 5 = slightly conservative, 6 = conservative, 7 = very conservative), PA = political affiliation (D = Democrat, R = Republican, I = independent), RE = how often you attend religious services (never, occasionally, most weeks, every week), LD = belief in life after death (yes, no), AA = support affirmative action (yes, no), AH = number of people you know who have died from AIDS or who are HIV+. You will use this data file for exercises in future chapters.

(a) Practice accessing a data file for statistical analysis with your software by going to this Web site and copying this data file. Print a copy of the data file. How many observations (rows) are in the data file?

(b) Give an example of a question that could be addressed using these data with (i) descriptive statistics, (ii) inferential statistics.

1.12. Using a spreadsheet program (such as Microsoft Office Excel) or statistical software, your instructor will help the class create a data file consisting of the values for class members of characteristics such as those in the previous exercise. One exercise in each chapter will use this data file.

(a) Copy the data file to your computer and print a copy.

(b) Give an example of a question that you could address by analyzing these data with (i) descriptive statistics, (ii) inferential statistics.

1.13. For the statistical software your instructor has chosen for your course, find out how to access the software, enter data, and print any data files that you create. Create a data file using the data in Figure 1.1 in Section 1.3, and print it.

1.14. Illustrating with an example, explain the difference between

(a) a *statistic* and a *parameter*.

(b) *description* and *inference* as two purposes for using statistical methods.

1.15. You have data for a population, from a census. Explain why descriptive statistics are helpful but inferential statistics are not needed.

1.16. A sociologist wants to estimate the average age at marriage for women in New England in the early eighteenth century. She finds within her state archives marriage records for a large Puritan village for the years 1700–1730. She then takes a sample of those records, noting the age of the bride for each. The average age in the sample is 24.1 years. Using a statistical method from Chapter 5, the sociologist estimates the average age of brides at marriage for the population to be between 23.5 and 24.7 years.

(a) What part of this example is descriptive?

(b) What part of this example is inferential?

(c) To what population does the inference refer?

1.17. In a recent survey by Eurobarometer of Europeans about energy issues and global warming,[1] one question asked, "Would you be willing to pay more for energy produced from renewable sources than for energy produced from other sources?" The percentage of *yes* responses varied among countries between 10% (in Bulgaria) to 60% (in Luxembourg). Of the 631 subjects interviewed in the UK, 45% said *yes*. It was predicted that for all 48 million adults in the UK, that percentage who would answer *yes* falls between 41% and 49%. Identify in this discussion (a) a statistic, (b) a parameter, (c) a descriptive statistical analysis, (d) an inferential statistical analysis.

1.18. Go to the Web site for the Gallup poll, www.galluppoll.com. From information listed on or linked from the homepage, give an example of a (a) descriptive statistical analysis, (b) inferential statistical analysis.

1.19. Check whether you have access to JSTOR (Journal Storage) at your school by visiting www.jstor.org. If so, click on *Browse* and then *Sociology* or another discipline of interest to you. Select a journal and a particular issue, and browse through some of the articles. Find an article that uses statistical methods. In a paragraph of 100–200 words, explain how descriptive statistics were used.

[1] *Attitudes towards Energy*, published January 2006 at ec.europa.eu/public_opinion

This page intentionally left blank

CHAPTER 2

Sampling and Measurement

To analyze social phenomena with a statistical analysis, *descriptive* methods summarize the data and *inferential* methods use sample data to make predictions about populations. In gathering data, we must decide which subjects to sample. Selecting a sample that is representative of the population is a primary topic of this chapter.

Given a sample, we must convert our ideas about social phenomena into data through deciding what to measure and how to measure it. Developing ways to measure abstract concepts such as achievement, intelligence, and prejudice is one of the most challenging aspects of social research. A measure should have *validity*, describing what it is intended to measure and accurately reflecting the concept. It should also have *reliability*, being consistent in the sense that a subject will give the same response when asked again. Invalid or unreliable data-gathering instruments render statistical manipulations of the data meaningless.

The first section of this chapter introduces definitions pertaining to measurement, such as types of data. The other sections discuss ways, good and bad, of selecting the sample.

2.1 VARIABLES AND THEIR MEASUREMENT

Statistical methods help us determine the factors that explain *variability* among subjects. For instance, variation occurs from student to student in their college grade point average (GPA). What is responsible for that variability? The way those students vary in how much they study per week? in how much they watch TV per day? in their IQ? in their college board score? in their high school GPA?

Variables

Any characteristic we can measure for each subject is called a ***variable***. The name reflects that values of the characteristic *vary* among subjects.

Variable

A ***variable*** is a characteristic that can vary in value among subjects in a sample or population.

Different subjects may have different values of a variable. Examples of variables are income last year, number of siblings, whether employed, and gender. The values the variable can take form the ***measurement scale***. For gender, for instance, the

measurement scale consists of the two labels, female and male. For number of siblings it is 0, 1, 2, 3, ...

The valid statistical methods for a variable depend on its measurement scale. We treat a numerical-valued variable such as annual income differently than a variable with a measurement scale consisting of categories, such as (yes, no) for whether employed. We next present ways to classify variables. The first type refers to whether the measurement scale consists of categories or numbers. Another type refers to the number of levels in that scale.

Quantitative and Categorical Variables

A variable is called ***quantitative*** when the measurement scale has numerical values. The values represent different magnitudes of the variable. Examples of quantitative variables are a subject's annual income, number of siblings, age, and number of years of education completed.

A variable is called ***categorical*** when the measurement scale is a set of categories. For example, marital status, with categories (single, married, divorced, widowed), is categorical. For Canadians, the province of residence is categorical, with the categories Alberta, British Columbia, and so on. Other categorical variables are whether employed (yes, no), primary clothes shopping destination (local mall, local downtown, Internet, other), favorite type of music (classical, country, folk, jazz, rock), religious affiliation (Protestant, Catholic, Jewish, Muslim, other, none), and political party preference.

For categorical variables, distinct categories differ in quality, not in numerical magnitude. Categorical variables are often called ***qualitative***. We distinguish between categorical and quantitative variables because different statistical methods apply to each type. Some methods apply to categorical variables and others apply to quantitative variables. For example, the *average* is a statistical summary for a quantitative variable, because it uses numerical values. It's possible to find the average for a quantitative variable such as income, but not for a categorical variable such as religious affiliation or favorite type of music.

Nominal, Ordinal, and Interval Scales of Measurement

For a quantitative variable, the possible numerical values are said to form an ***interval*** scale. Interval scales have a specific numerical distance or *interval* between each pair of levels. Annual income is usually measured on an interval scale. The interval between $40,000 and $30,000, for instance, equals $10,000. We can compare outcomes in terms of how much larger or how much smaller one is than the other.

Categorical variables have two types of scales. For the categorical variables mentioned in the previous subsection, the categories are unordered. The scale does not have a "high" or "low" end. The categories are then said to form a ***nominal scale***. For another example, a variable measuring primary mode of transportation to work might use the nominal scale with categories (automobile, bus, subway, bicycle, walk).

Although the different categories are often called the *levels* of the scale, for a nominal variable no level is greater than or smaller than any other level. Names or labels such as "automobile" and "bus" for mode of transportation identify the categories but do not represent different magnitudes. By contrast, each possible value of a quantitative variable is *greater than* or *less than* any other possible value.

A third type of scale falls, in a sense, between nominal and interval. It consists of categorical scales having a natural *ordering* of values. The levels form an ***ordinal scale***. Examples are social class (upper, middle, lower), political philosophy (very liberal, slightly liberal, moderate, slightly conservative, very conservative),

government spending on the environment (too little, about right, too much), and frequency of religious activity (never, less than once a month, about 1–3 times a month, every week, more than once a week). These scales are not nominal, because the categories are ordered. They are not interval, because there is no defined distance between levels. For example, a person categorized as very conservative is *more* conservative than a person categorized as slightly conservative, but there is no numerical value for *how much more* conservative that person is.

In summary, for ordinal variables the categories have a natural ordering, whereas for nominal variables the categories are unordered. The scales refer to the actual measurement and not to the phenomena themselves. *Place of residence* may indicate a geographic place name such as a county (nominal), the distance of that place from a point on the globe (interval), the size of the place (interval or ordinal), or other kinds of variables.

Quantitative Aspects of Ordinal Data

As we've discussed, levels of nominal scales are qualitative, varying in quality, not in quantity. Levels of interval scales are quantitative, varying in magnitude. The position of ordinal scales on the quantitative–qualitative classification is fuzzy. Because their scale is a set of categories, they are often analyzed using the same methods as nominal scales. But in many respects, ordinal scales more closely resemble interval scales. They possess an important quantitative feature: Each level has a *greater* or *smaller* magnitude than another level.

Some statistical methods apply specifically to ordinal variables. Often, though, it's helpful to analyze ordinal scales by assigning numerical scores to categories. By treating ordinal variables as interval rather than nominal, we can use the more powerful methods available for quantitative variables.

For example, course grades (such as A, B, C, D, E) are ordinal. But we treat them as interval when we assign numbers to the grades (such as 4, 3, 2, 1, 0) to compute a grade point average. Treating ordinal variables as interval requires good judgment in assigning scores. In doing this, you can conduct a "sensitivity analysis" by checking whether conclusions would differ in any significant way for other choices of the scores.

Discrete and Continuous Variables

One other way to classify a variable also helps determine which statistical methods are appropriate for it. This classification refers to the number of values in the measurement scale.

Discrete and Continuous Variables

A variable is ***discrete*** if its possible values form a set of separate numbers, such as 0, 1, 2, 3, ... It is ***continuous*** if it can take an infinite continuum of possible real number values.

Examples of discrete variables are the number of siblings and the number of visits to a physician last year. Any variable phrased as "the number of . . . " is discrete, because it is possible to list its possible values $\{0, 1, 2, 3, 4, \ldots\}$.

Examples of continuous variables are height, weight, and the amount of time it takes to read a passage of a book. It is impossible to write down all the distinct potential values, since they form an interval of infinitely many values. The amount of time needed to read a book, for example, could take on the value 8.6294473. . . hours.

Discrete variables have a basic unit of measurement that cannot be subdivided. For example, 2 and 3 are possible values for the number of siblings, but 2.5716 is

not. For a continuous variable, by contrast, between any two possible values there is always another possible value. For example, age is continuous in the sense that an individual does not age in discrete jumps. At some well-defined point during the year in which you age from 21 to 22, you are 21.3851 years old, and similarly for every other real number between 21 and 22. A continuous, infinite collection of age values occurs between 21 and 22 alone.

Any variable with a finite number of possible values is discrete. All categorical variables, nominal or ordinal, are discrete, having a finite set of categories. Quantitative variables can be discrete or continuous; age is continuous, and number of siblings is discrete.

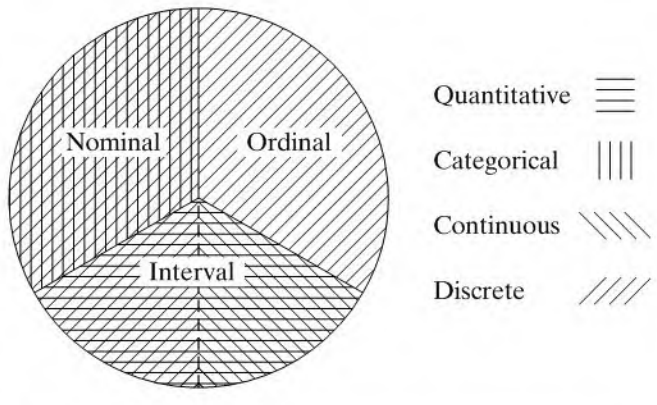
For quantitative variables the distinction between discrete and continuous variables can be blurry, because of how variables are actually measured. In practice, we round continuous variables when measuring them, so the measurement is actually discrete. We say that an individual is 21 years old whenever that person's age is somewhere between 21 and 22. On the other hand, some variables, although discrete, have a very large number of possible values. In measuring annual family income in dollars, the potential values are 0, 1, 2, 3, ..., up to some very large value in many millions.

What's the implication of this? Statistical methods for discrete variables are mainly used for quantitative variables that take relatively few values, such as the number of times a person has been married. Statistical methods for continuous variables are used for quantitative variables that can take lots of values, regardless of whether they are theoretically continuous or discrete. For example, statisticians treat variables such as age, income, and IQ as continuous.

In summary,

- Variables are either *quantitative* (numerical valued) or *categorical*. Quantitative variables are measured on an *interval* scale. Categorical variables with unordered categories have a *nominal* scale, and categorical variables with ordered categories have an *ordinal* scale.
- Categorical variables (nominal or ordinal) are *discrete*. Quantitative variables can be either discrete or continuous. In practice, quantitative variables that can take lots of values are treated as *continuous*.

Figure 2.1 summarizes the types of variables, in terms of the (quantitative, categorical), (nominal, ordinal, interval), and (continuous, discrete) classifications.



Note: Ordinal data are treated sometimes as categorical and sometimes as quantitative

FIGURE 2.1: Summary of Quantitative–Categorical, Nominal–Ordinal–Interval, Continuous–Discrete Classifications

2.2 RANDOMIZATION

Inferential statistical methods use sample statistics to make predictions about population parameters. The quality of the inferences depends on how well the sample represents the population. This section introduces an important sampling method that incorporates ***randomization***, the mechanism for achieving good sample representation.

Simple Random Sampling

Subjects of a population to be sampled could be individuals, families, schools, cities, hospitals, records of reported crimes, and so on. *Simple random sampling* is a method of sampling for which every possible sample has equal chance of selection.

Let n denote the number of subjects in the sample, called the ***sample size***.

Simple Random Sample

A ***simple random sample*** of n subjects from a population is one in which each possible sample of that size has the same probability (chance) of being selected.

For instance, suppose a researcher administers a questionnaire to one randomly selected adult in each of several households. A particular household contains four adults—mother, father, aunt, and uncle—identified as M, F, A, and U. For a simple random sample of $n = 1$ adult, each of the four adults is equally likely to be interviewed. You could select one by placing the four names on four identical ballots and selecting one blindly from a hat. For a simple random sample of $n = 2$ adults, each possible sample of size two is equally likely. The six potential samples are (M, F), (M, A), (M, U), (F, A), (F, U), and (A, U). To select the sample, you blindly select two ballots from the hat.

A simple random sample is often just called a ***random sample***. The *simple* adjective is used to distinguish this type of sampling from more complex sampling schemes presented in Section 2.4 that also have elements of randomization.

Why is it a good idea to use random sampling? Because everyone has the same chance of inclusion in the sample, so it provides fairness. This reduces the chance that the sample is seriously biased in some way, leading to inaccurate inferences about the population. Most inferential statistical methods assume randomization of the sort provided by random sampling.

How to Select a Simple Random Sample

To select a random sample, we need a list of all subjects in the population. This list is called the ***sampling frame***. Suppose you plan to sample students at your school. The population is all students at the school. One possible sampling frame is the student directory.

The most common method for selecting a random sample is to (1) number the subjects in the sampling frame, (2) generate a set of these numbers randomly, and (3) sample the subjects whose numbers were generated. Using *random numbers* to select the sample ensures that each subject has an equal chance of selection.

Random Numbers

Random numbers are numbers that are computer generated according to a scheme whereby each digit is equally likely to be any of the integers $0, 1, 2, \ldots, 9$ and does not depend on the other digits generated.

TABLE 2.1: Part of a Table of Random Numbers

Line/Col.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	10480	15011	01536	02011	81647	91646	69179	14194
2	22368	46573	25595	85393	30995	89198	27982	53402
3	24130	48360	22527	97265	76393	64809	15179	24830
4	42167	93093	06243	61680	07856	16376	39440	53537
5	37570	39975	81837	16656	06121	91782	60468	81305
6	77921	06907	11008	42751	27756	53498	18602	70659

Source: Abridged from William H. Beyer, ed., *Handbook of Tables for Probability and Statistics*, 2nd ed., © The Chemical Rubber Co., 1968. Used by permission of the Chemical Rubber Co.

Table 2.1 shows a table containing random numbers. The numbers fluctuate according to no set pattern. Any particular number has the same chance of being a 0, 1, 2, ..., or 9. The numbers are chosen independently, so any one digit chosen has no influence on any other selection. If the first digit in a row of the table is a 9, for instance, the next digit is still just as likely to be a 9 as a 0 or 1 or any other number. Random numbers are available in published tables and can be generated with software and many statistical calculators.

Suppose you want to select a simple random sample of $n = 100$ students from a university student body of size 30,000. The sampling frame is a directory of these students. Select the students by using five-digit sequences to identify them, as follows:

1. Assign the numbers 00001 to 30000 to the students in the directory, using 00001 for the first student in the list, 00002 for the second student, and so on.
2. Starting at any point in the random number table, or by generating random numbers using software or a calculator, choose successive five-digit numbers until you obtain 100 distinct numbers between 00001 and 30000.
3. Include in the sample the students with assigned numbers equal to the random numbers selected.

For example, for the first column of five-digit numbers in Table 2.1, the first three random numbers are 10480, 22368, and 24130. The first three students selected are those numbered 10480, 22368, and 24130.

In selecting the 100 five-digit numbers, skip numbers greater than 30000, such as the next three five-digit numbers in Table 2.1, since no student in the directory has an assigned number that large. After using the first column of five-digit numbers, move to the next column of numbers and continue. If the population size were between 1000 and 9999, you would use four digits at a time. The column (or row) from which you begin selecting the numbers does not matter, since the numbers have no set pattern. Most statistical software can do this all for you.

Collecting Data with Sample Surveys

Many studies select a sample of people from a population and interview them to collect data. This method of data collection is called a ***sample survey***. The interview could be a personal interview, telephone interview, or self-administered questionnaire.

The General Social Survey (GSS) is an example of a sample survey. It gathers information using personal interviews of a random sample of subjects from the U.S. adult population to provide a snapshot of that population. (The survey does not use *simple* random sampling but rather a method discussed later in the chapter that

incorporates multiple stages and clustering but is designed to give each family the same chance of inclusion.) National polls such as the Gallup Poll are also sample surveys. They usually use telephone interviews. Since it is often difficult to obtain a sampling frame, many telephone interviews obtain the sample with *random digit dialing*.

A variety of problems can cause responses from a sample survey to tend to favor some parts of the population over others. Then results from the sample are not representative of the population. We'll discuss some potential problems in Section 2.3.

Collecting Data with an Experiment

In some studies, data result from a planned *experiment*. The purpose of most experiments is to compare responses of subjects on some outcome measure, under different conditions. Those conditions are levels of a variable that can influence the outcome. The scientist has the experimental control of being able to assign subjects to the conditions.

For instance, the conditions might be different drugs for treating some illness. The conditions are called ***treatments***. To conduct the experiment, the researcher needs a plan for assigning subjects to the treatments. These plans are called ***experimental designs***. Good experimental designs use randomization to determine which treatment a subject receives.

In the late 1980s, the Physicians' Health Study Research Group at Harvard Medical School designed an experiment to analyze whether regular intake of aspirin reduces mortality from heart disease. Of about 22,000 male physicians, half were randomly chosen to take an aspirin every other day. The remaining half took a placebo, which had no active agent. After five years, rates of heart attack were compared. By using randomization to determine who received which treatment, the researchers knew the groups would roughly balance on factors that could affect heart attack rates, such as age and quality of health. If the physicians could decide on their own which treatment to take, the groups might have been out of balance on some important factor. Suppose, for instance, younger physicians were more likely to select aspirin. Then, a lower heart attack rate among the aspirin group could occur merely because younger subjects are less likely to suffer heart attacks.

Collecting Data with an Observational Study

In social research, it is rarely possible to conduct experiments. It's not possible to randomly assign subjects to the groups we want to compare, such as levels of gender or race or educational level or annual income. Many studies merely *observe* the outcomes for available subjects on the variables without any experimental manipulation of the subjects. Such studies are called ***observational studies***. The researcher measures subjects' responses on the variables of interest but has no experimental control over the subjects.

With observational studies, comparing groups is difficult because the groups may be imbalanced on variables that affect the outcome. This is true even with random sampling. For instance, suppose we plan to compare black students, Hispanic students, and white students on some standardized test. If white students have a higher average score, a variety of variables might account for that difference, such as parents' education or parents' income or quality of school attended. This makes it difficult to compare groups with observational studies, especially when some key variables may not have been measured in the study.

Establishing cause and effect is central to science. But it's not possible to establish cause and effect definitively with a nonexperimental study, whether it be an observational study with an available sample or a sample survey using random sampling. With an observational study, there's the strong possibility that the sample does not well reflect the population. With an observational study or a sample survey, there's always the possibility that some unmeasured variable could be responsible for patterns observed in the data. With an experiment that randomly assigns subjects to treatments, those treatments should roughly balance on any unmeasured variables. For example, in the heart attack study mentioned above, the doctors taking aspirin would not tend to be younger or of better health than the doctors taking placebo. Because a randomized experiment balances the groups being compared on other factors, it's possible to study cause and effect more accurately with an experiment than with an observational study.

Whether or not a study is experimental, it's important to incorporate randomization in any study that plans to make inferences. This randomization could take the form of randomly selecting a sample for a survey, or randomly allocating subjects to different treatments for an experimental study.

2.3 SAMPLING VARIABILITY AND POTENTIAL BIAS

Even if a study wisely uses randomization, the results of the study still depend on which subjects are sampled. Two researchers who separately select random samples from some population may have little overlap, if any, between the two sample memberships. Therefore, the values of sample statistics will differ for the two samples, and the results of analyses based on these samples may differ.

Sampling Error

Suppose the Gallup, Harris, Zogby, and Pew polling organizations each randomly sample 1000 adult Canadians, in order to estimate the percentage of Canadians who give the prime minister's performance in office a favorable rating. Based on the samples they select, perhaps Gallup reports an approval rating of 63%, Harris reports 68%, Zogby 65%, and Pew 64%. These differences could reflect slightly different question wording. But even if the questions are worded exactly the same, the percentages would probably differ somewhat because the samples are different.

For conclusions based on statistical inference to be worthwhile, we should know the potential *sampling error*—how much the statistic differs from the parameter it predicts because of the way results naturally exhibit variation from sample to sample.

Sampling Error

The ***sampling error*** of a statistic equals the error that occurs when we use a statistic based on a sample to predict the value of a population parameter.

Suppose that the actual percentage of the population of adult Canadians who give the prime minister a favorable rating is 66%. Then the Gallup organization, which predicted 63%, had a sampling error of $63\% - 66\% = -3\%$. The Harris organization, which predicted 68%, had a sampling error of $68\% - 66\% = 2\%$. In practice, the sampling error is unknown, because the values of population parameters are unknown.

Random sampling protects against bias, in the sense that the sampling error tends to fluctuate about 0, sometimes being positive (as in the Harris poll) and sometimes being negative (as in the Gallup poll). Random sampling also allows us to predict the likely size of the sampling error. For sample sizes of about 1000, we'll see that

the sampling error for estimating percentages is usually no greater than plus or minus 3%. This bound is the *margin of error*. Variability also occurs in the values of sample statistics with nonrandom sampling, but the extent of the sampling error is not predictable as it is with random samples.

Sampling Bias: Nonprobability Sampling

Other factors besides sampling error can cause results to vary from sample to sample. These factors can also possibly cause bias. We next discuss three types of bias. The first is called ***sampling bias***.

For simple random sampling, each possible sample of n subjects has the same probability of selection. This is a type of ***probability sampling*** method, meaning that the probability any particular sample will be selected is known. Inferential statistical methods assume probability sampling. ***Nonprobability sampling*** methods are ones for which it is not possible to determine the probabilities of the possible samples. Inferences using such samples have unknown reliability and result in ***sampling bias***.

The most common nonprobability sampling method is ***volunteer sampling***. As the name implies, subjects volunteer to be in the sample. But the sample may poorly represent the population and yield misleading conclusions. For instance, a mail-in questionnaire published in *TV Guide* posed the question, "Should the President have the Line Item Veto to eliminate waste?" Of those who responded, 97% said yes. For the same question posed to a random sample, 71% said yes.[1]

Examples of volunteer sampling are visible any day on many Internet sites and television news programs. Viewers register their opinions on an issue by voting over the Internet. The viewers who respond are unlikely to be a representative cross section, but will be those who can easily access the Internet and who feel strongly enough to respond. Individuals having a particular opinion might be much more likely to respond than individuals having a different opinion. For example, one night the ABC program *Nightline* asked viewers whether the United Nations should continue to be located in the United States. Of more than 186,000 respondents, 67% wanted the United Nations out of the United States. At the same time, a poll using a random sample of about 500 respondents estimated the population percentage to be about 28%. Even though the random sample had a much smaller size, it is far more trustworthy.

A large sample does not help with volunteer sampling—the bias remains. In 1936, the newsweekly *Literary Digest* sent over 10 million questionnaires in the mail to predict the outcome of the presidential election. The questionnaires went to a relatively wealthy segment of society (those having autos or telephones), and fewer than 25% were returned. The journal used these to predict an overwhelming victory by Alfred Landon over Franklin Roosevelt. The opposite result was predicted by George Gallup with a much smaller sample in the first scientific poll taken for this purpose. In fact, Roosevelt won in a landslide.

Unfortunately, volunteer sampling is sometimes necessary. This is often true in medical studies. Suppose a study plans to investigate how well a new drug performs compared to a standard drug, for subjects who suffer from high blood pressure. The researchers are not going to be able to find a sampling frame of all who suffer from high blood pressure and take a simple random sample of them. They may, however, be able to sample such subjects at certain medical centers or using volunteers. Even then, randomization should be used wherever possible. For the study patients, they can randomly select who receives the new drug and who receives the standard one.

---

[1]D. M. Wilbur, *Public Perspective*, available at roperweb.ropercenter.uconn.edu, May–June 1993.

Even with random sampling, sampling bias can occur. One case is when the sampling frame suffers from ***undercoverage***: It lacks representation from some groups in the population. A telephone survey will not reach prison inmates or homeless people or people too poor to afford a telephone, whereas families that have many phones will tend to be over-represented. Responses by those not having a telephone might tend to be quite different from those actually sampled, leading to biased results.

Response Bias

In a survey, the way a question is worded or asked can have a large impact on the results. For example, when a *New York Times*/CBS News poll in 2006 asked whether the interviewee would be in favor of a new gasoline tax, only 12% said yes. When the tax was presented as reducing U.S. dependence on foreign oil, 55% said yes, and when asked about a gas tax that would help reduce global warming, 59% said yes.[2]

Poorly worded or confusing questions result in ***response bias***. Even the order in which questions are asked can influence the results dramatically. During the Cold War, a study asked, "Do you think the U.S. should let Russian newspaper reporters come here and send back whatever they want?" and "Do you think Russia should let American newspaper reporters come in and send back whatever they want?" The percentage of yes responses to the first question was 36% when it was asked first and 73% when it was asked second.[3]

In an interview, characteristics of the interviewer may result in response bias. Respondents might lie if they think their belief is socially unacceptable. They may be more likely to give the answer that they think the interviewer prefers. An example is provided by a study on the effect of the interviewer's race. Following a phone interview, respondents were asked whether they thought the interviewer was black or white (all were actually black). Perceiving a white interviewer resulted in more conservative opinions. For example, 14% agreed that "American society is fair to everyone" when they thought the interviewer was black, but 31% agreed to this when they thought the interviewer was white.[4]

Nonresponse Bias: Missing Data

Some subjects who are supposed to be in the sample may refuse to participate, or it may not be possible to reach them. This results in the problem of ***nonresponse bias***. If only half the intended sample was actually observed, we should worry about whether the half not observed differ from those observed in a way that causes biased results. Even if we select the sample randomly, the results are questionnable if there is substantial nonresponse, say, over 20%.

For her book *Women in Love*, author Shere Hite surveyed women in the United States. One of her conclusions was that 70% of women who had been married at least five years have extramarital affairs. She based this conclusion on responses to questionnaires returned by 4500 women. This sounds like an impressively large sample. However, the questionnaire was mailed to about 100,000 women. We cannot know whether the 4.5% of the women who responded were representative of the 100,000 who received the questionnaire, much less the entire population of American women. This makes it dangerous to make an inference to the larger population.

---

[2]Column by T. Friedman, *New York Times*, March 2, 2006.

[3]See Crosson (1994).

[4]*Washington Post*, June 26, 1995.

Missing data is a problem in almost all large studies. Some subjects do not provide responses for some of the variables measured. Even in censuses, which are designed to observe everyone in a country, some people are not observed or fail to cooperate. Most software ignores cases for which observations are missing for at least one variable used in an analysis. This results in wasted information and possible bias. Statisticians have recently developed methods that replace missing observations by predicted values based on patterns in the data. See Allison (2002) for an introduction to ways of dealing with missing data.

Summary of Types of Bias

In summary, sample surveys have potential sources of bias:

- ***Sampling bias*** occurs from using nonprobability samples or having undercoverage.
- ***Response bias*** occurs when the subject gives an incorrect response (perhaps lying), or the question wording or the way the interviewer asks the questions is confusing or misleading.
- ***Nonresponse bias*** occurs when some sampled subjects cannot be reached or refuse to participate or fail to answer some questions.

In any study, carefully assess the scope of conclusions. Evaluate critically the conclusions by noting the makeup of the sample. How was the sample selected? How large was it? How were the questions worded? Who sponsored and conducted the research? The less information that is available, the less you should trust it.

Finally, be wary of any study that makes inferences to a broader population than is justified by the sample chosen. Suppose a psychologist performs an experiment using a random sample of students from an introductory psychology course. With statistical inference, the sample results generalize to the population of all students in the class. For the results to be of wider interest, the psychologist might claim that the conclusions extend to *all* college students, to all young adults, or even to all adults. These generalizations may well be wrong, because the sample may differ from those populations in fundamental ways, such as in average age or socioeconomic status.

2.4 OTHER PROBABILITY SAMPLING METHODS*

Section 2.2 introduced ***simple random sampling*** and explained its importance to statistical inference. In practice, other probability sampling methods that have elements of randomness are sometimes preferable to simple random sampling or are simpler to obtain.

Systematic Random Sampling

Systematic random sampling selects a subject near the beginning of the sampling frame list, skips several names and selects another subject, skips several more names and selects the next subject, and so forth. The number of names skipped at each stage depends on the desired sample size. Here's how it's done:

Systematic Random Sample

Denote the sample size by n and the population size by N. Let $k = N/n$, the population size divided by the sample size. A ***systematic random sample*** (1) selects a subject at random from the first k names in the sampling frame, and (2) selects every kth subject listed after that one. The number k is called the *skip number*.

Suppose you want a systematic random sample of 100 students from a population of 30,000 students listed in a campus directory. Then, $n = 100$ and $N = 30{,}000$, so $k = 30{,}000/100 = 300$. The population size is 300 times the sample size, so you need to select one of every 300 students. You select one student at random, using random numbers, from the first 300 students in the directory. Then you select every 300th student after the one selected randomly. This produces a sample of size 100. The first three digits in Table 2.1 are 104, which falls between 001 and 300, so you first select the student numbered 104. The numbers of the other students selected are $104 + 300 = 404$, $404 + 300 = 704$, $704 + 300 = 1004$, $1004 + 300 = 1304$, and so on. The 100th student selected is listed in the last 300 names in the directory.

In sampling from a sampling frame, it's simpler to select a systematic random sample than a simple random sample because it uses only one random number. This method typically provides as good a representation of the population, because for alphabetic listings such as directories of names, values of most variables fluctuate randomly through the list. With this method, statistical formulas based on simple random sampling are usually valid.

A systematic random sample is not a simple random sample, because all samples of size n are not equally likely. For instance, unlike in a simple random sample, two subjects listed next to each other on the list cannot both appear in the sample.

Stratified Random Sampling

Another probability sampling method, useful in social science research for studies comparing groups, is *stratified sampling*.

Stratified Random Sample

A ***stratified random sample*** divides the population into separate groups, called ***strata***, and then selects a simple random sample from each stratum.

Suppose a study in Cambridge, Massachusetts plans to compare the opinions of registered Democrats and registered Republicans about whether government should guarantee health care to all citizens. Stratifying according to political party registration, the study selects a random sample of Democrats and another random sample of Republicans.

Stratified random sampling is called ***proportional*** if the sampled strata proportions are the same as those in the entire population. For example, if 90% of the population of interest is Democrat and 10% is Republican, then the sampling is proportional if the sample size for Democrats is nine times the sample size for Republicans.

Stratified random sampling is called ***disproportional*** if the sampled strata proportions differ from the population proportions. This is useful when the population size for a stratum is relatively small. A group that comprises a small part of the population may not have enough representation in a simple random sample to allow precise inferences. It is not possible to compare accurately Democrats to Republicans, for example, if only 10 people in a sample size of 100 are Republican. By contrast, a disproportional stratified sample size of 100 might randomly sample 50 Democrats and 50 Republicans.

To implement stratification, we must know the stratum into which each subject in the sampling frame belongs. This usually restricts the variables that can be used for forming the strata. The variables must have strata that are easily identifiable. For example, it would be easy to select a stratified sample of a school population

using grade level as the stratification variable, but it would be difficult to prepare an adequate sampling frame of city households stratified by household income.

Cluster Sampling

Simple, systematic, and stratified random sampling are often difficult to implement, because they require a complete sampling frame. Such lists are easy to obtain for sampling cities or hospitals or schools, for example, but more difficult for sampling individuals or families. *Cluster sampling* is useful when a complete listing of the population is not available.

Cluster Random Sample

Divide the population into a large number of ***clusters***, such as city blocks. Select a simple random sample of the clusters. Use the subjects in those clusters as the sample.

For example, a study might plan to sample about 1% of the families in a city, using city blocks as clusters. Using a map to identify city blocks, it could select a simple random sample of 1% of the blocks and then sample every family on each block. A study of patient care in mental hospitals in Ontario could first randomly sample mental hospitals (the clusters) and then collect data for patients within those hospitals.

What's the difference between a stratified sample and a cluster sample? A stratified sample uses *every* stratum. The strata are usually groups we want to compare. By contrast, a cluster sample uses a *sample* of the clusters, rather than all of them. In cluster sampling, clusters are merely ways of easily identifying groups of subjects. The goal is not to compare the clusters but rather to use them to obtain a sample. Most clusters are not represented in the eventual sample.

Figure 2.2 illustrates the distinction among sampling subjects (simple random sample), sampling clusters of subjects (cluster random sample), and sampling subjects from within strata (stratified random sample). The figure depicts ways to survey 40 students at a school, to make comparisons among Freshmen, Sophomores, Juniors, and Seniors.

Multistage Sampling

When conducting a survey for predicting elections, the Gallup Organization often identifies election districts as clusters and takes a simple random sample of them. But then it also takes a simple random sample of households within each selected election

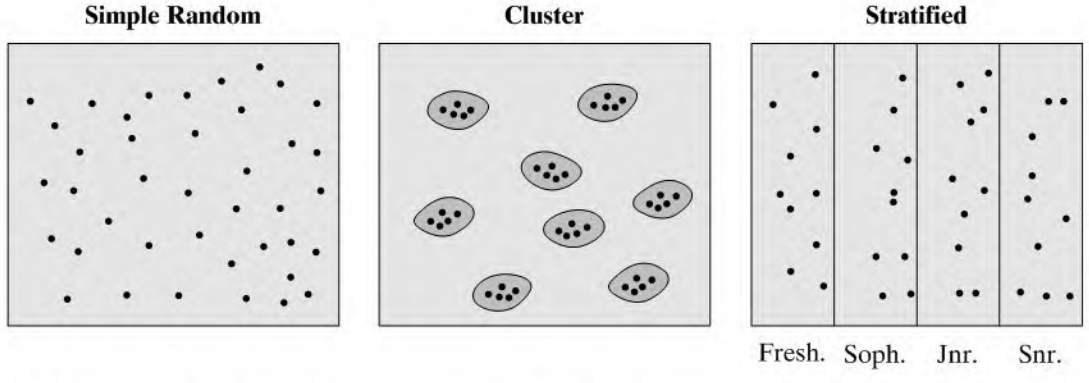


FIGURE 2.2: Ways of Randomly Sampling 40 Students. The figure is a schematic for a simple random sample, a cluster random sample of 8 clusters of students who live together, and a stratified random sample of 10 students from each class (Fr, So, Ju, Sr).

district. This is more feasible than sampling *every* household in the chosen districts. This is an example of *multistage sampling*, which uses combinations of sampling methods.

Here's an example of a multistage sample:

- Treat counties (or census tracts) as clusters and select a random sample of a certain number of them.
- Within each county selected, take a cluster random sample of square-block regions.
- Within each region selected, take a systematic random sample of every tenth house.
- Within each house selected, select one adult at random for the sample.

Multistage samples are common in social science research. They are simpler to implement than simple random sampling but provide a broader sampling of the population than a single method such as cluster sampling.

For statistical inference, stratified samples, cluster samples, and multistage samples use different formulas from the ones in this book. Cluster sampling requires a larger sample to achieve as much inferential precision as simple random sampling. Observations within clusters tend to be similar, because of the tendency of subjects living near one another to have similar values on opinion issues and on economic and demographic variables such as age, income, race, and occupation. So we need more data to obtain a representative cross section. By contrast, the results for stratified sampling may be more precise than those stated in this textbook for simple random sampling. Books specializing in sampling methodology provide further details (Scheaffer et al., 2005; Thompson, 2002).

2.5 CHAPTER SUMMARY

Statistical methods analyze data on ***variables***, which are characteristics that vary among subjects. The statistical methods used depend on the type of variable:

- Numerically measured variables, such as family income and number of children in a family, are ***quantitative***. They are measured on an *interval scale*.
- Variables taking value in a set of categories are ***categorical***. Those measured with unordered categories, such as religious affiliation and province of residence, have a *nominal scale*. Those measured with ordered categories, such as social class and political ideology, have an *ordinal scale* of measurement.
- Variables are also classified as ***discrete***, having possible values that are a set of separate numbers (such as 0, 1, 2, ...), or ***continuous***, having a continuous, infinite set of possible values. Categorical variables, whether nominal or ordinal, are discrete. Quantitative variables can be of either type, but in practice are treated as continuous if they can take a large number of values.

Much social science research uses ***observational studies***, which use available subjects to observe variables of interest. One should be cautious in attempting to conduct inferential analyses with data from such studies. Inferential statistical methods require ***probability samples***, which incorporate randomization in some way. Random sampling allows control over the amount of ***sampling error***, which describes how results can vary from sample to sample. Random samples are much more likely to be representative of the population than are nonprobability samples such as volunteer samples.

- For a ***simple random sample***, every possible sample of size n has the same chance of selection.

- Here are other examples of probability sampling: ***Systematic*** random sampling takes every kth subject in the sampling frame list. ***Stratified*** random sampling divides the population into groups (strata) and takes a random sample from each stratum. ***Cluster*** random sampling takes a random sample of clusters of subjects (such as city blocks) and uses subjects in those clusters as the sample. ***Multistage*** sampling uses combinations of these methods.

Chapter 3 introduces statistics for describing samples and corresponding parameters for describing populations. Hence, its focus is on *descriptive statistics*.

PROBLEMS

Practicing the Basics

2.1. Explain the difference between
- **(a)** Discrete and continuous variables
- **(b)** Categorical and quantitative variables
- **(c)** Nominal and ordinal variables

Why do these distinctions matter for statistical analysis?

2.2. Identify each variable as categorical or quantitative:
- **(a)** Number of pets in family
- **(b)** County of residence
- **(c)** Choice of auto (domestic or import)
- **(d)** Distance (in miles) commute to work
- **(e)** Choice of diet (vegetarian, nonvegetarian)
- **(f)** Time spent in previous month browsing the World Wide Web
- **(g)** Ownership of personal computer (yes, no)
- **(h)** Number of people you have known with AIDS (0, 1, 2, 3, 4 or more)
- **(i)** Marriage form of a society (monogamy, polygyny, polyandry)

2.3. Which scale of measurement (nominal, ordinal, or interval) is most appropriate for
- **(a)** Attitude toward legalization of marijuana (favor, neutral, oppose)
- **(b)** Gender (male, female)
- **(c)** Number of children in family (0, 1, 2, ...)
- **(d)** Political party affiliation (Democrat, Republican, Independent)
- **(e)** Religious affiliation (Catholic, Jewish, Protestant, Muslim, other)
- **(f)** Political philosophy (very liberal, somewhat liberal, moderate, somewhat conservative, very conservative)
- **(g)** Years of school completed (0, 1, 2, 3, ...)
- **(h)** Highest degree attained (none, high school, bachelor's, master's, doctorate)
- **(i)** College major (education, anthropology, physics, sociology, ...)
- **(j)** Test score (0–100 range for scores)
- **(k)** Employment status (employed full time, employed part time, unemployed)

2.4. Which scale of measurement is most appropriate for
- **(a)** Occupation (plumber, teacher, secretary, ...)
- **(b)** Occupational status (blue collar, white collar)
- **(c)** Social status (lower, middle, upper class)
- **(d)** Statewide murder rate (number of murders per 1000 population)
- **(e)** County population size (number of people)
- **(f)** Population growth rate (in percentages)
- **(g)** Community size (rural, small town, large town, small city, large city)
- **(h)** Annual income (thousands of dollars per year)
- **(i)** Attitude toward affirmative action (favorable, neutral, unfavorable)
- **(j)** Lifetime number of sexual partners

2.5. Which scale of measurement is most appropriate for "attained education" measured as
- **(a)** Number of years (0, 1, 2, 3, ...)
- **(b)** Grade level (elementary school, middle school, high school, college, graduate school)
- **(c)** School type (public school, private school)

2.6. Give an example of a variable that is **(a)** categorical, **(b)** quantitative, **(c)** ordinal scale, **(d)** nominal scale, **(e)** discrete, **(f)** continuous, **(g)** quantitative and discrete.

2.7. A poll conducted by YouGov for the British newspaper *The Daily Telegraph* in June 2006 asked a random sample of 1962 British adults several questions about their image of the U.S. One question asked, "How would you rate George W. Bush as a world leader?" The possible choices were (great leader, reasonably satisfactory leader, pretty poor leader, terrible leader).
- **(a)** Is this four-category variable nominal, or ordinal? Why?
- **(b)** Is this variable continuous, or discrete? Why?
- **(c)** Of the 93% of the sample who responded, the percentages in the four categories were 1% (great leader), 16%, 37%, 46% (terrible leader). Are these values statistics, or parameters? Why?

2.8. A survey asks subjects to rate five issues according to their importance in determining voting intention

for U.S. senator, using the scale (very important, somewhat important, unimportant). The issues are foreign policy, unemployment, inflation, the arms race, and civil rights. The evaluations can be treated as five variables: foreign policy evaluation, unemployment evaluation, and so on. These variables represent what scale of measurement?

2.9. Which of the following variables could theoretically be measured on a continuous scale? **(a)** Method of contraception used, **(b)** length of time of residence in a state, **(c)** task completion time, **(d)** intelligence, **(e)** authoritarianism, **(f)** alienation, **(g)** county of residence.

2.10. Which of the following variables are continuous when the measurements are as fine as possible? **(a)** Age of mother, **(b)** number of children in family, **(c)** income of spouse, **(d)** population of cities, **(e)** latitude and longitude of cities, **(f)** distance of home from place of employment, **(g)** number of foreign languages spoken.

2.11. A class has 50 students. Use the column of the first two digits in the random number table (Table 2.1) to select a simple random sample of three students. If the students are numbered 01 to 50, what are the numbers of the three students selected?

2.12. A local telephone directory has 400 pages with 130 names per page, a total of 52,000 names. Explain how you could choose a simple random sample of 5 names. Using the second column of Table 2.1 or software or a calculator, select 5 random numbers to identify subjects for the sample.

2.13. Explain whether an experiment or an observational study would be more appropriate to investigate the following:

(a) Whether or not cities with higher unemployment rates tend to have higher crime rates

(b) Whether a Honda Accord or a Toyota Camry gets better gas mileage

(c) Whether or not higher college GPAs tend to occur for students who had higher scores on college entrance exams

(d) Whether or not a special coupon attached to the outside of a catalog makes recipients more likely to order products from a mail-order company

2.14. A study is planned to study whether passive smoking (being exposed to secondhand cigarette smoke on a regular basis) leads to higher rates of lung cancer.

(a) One possible study is to take a sample of children, randomly select half of them for placement in an environment where they are passive smokers, and place the other half in an environment where they are not exposed to smoke. Then 60 years later the observation is whether each has developed lung cancer. Would this study be an experimental study or an observational study? Why?

(b) For many reasons, including time and ethics, it is not possible to conduct the study in (a). Describe a way that *is* possible, and indicate whether it would be an experimental or observational study.

2.15. Table 2.2 shows the result of the 2000 Presidential election and the predictions of several organizations in the days before the election. The sample sizes were typically about 2000. The percentages for each poll do not sum to 100 because of voters reporting as undecided or favoring another candidate.

(a) What factors cause the results to vary somewhat among organizations?

(b) Identify the sampling error for the Gallup poll.

TABLE 2.2

	Predicted Vote		
Poll	Gore	Bush	Nader
Gallup	46	48	4
Harris	47	47	5
ABC	45	48	3
CBS	45	44	4
NBC	44	47	3
Pew Research	47	49	4
Actual vote	48.4	47.9	2.7

Source: www.ncpp.org/

2.16. The BBC in Britain requested viewers to call the network and indicate their favorite poem. Of more than 7500 callers, more than twice as many voted for Rudyard Kipling's *If* than for any other poem. The BBC reported that this was the clear favorite.

(a) Explain what it means to call this a "volunteer sample."

(b) If the BBC truly wanted to determine Brits' favorite poem, how could it more reliably do so?

2.17. A Roper Poll was designed to determine the percentage of Americans who express some doubt that the Nazi Holocaust occurred. In response to the question, "Does it seem possible or does it seem impossible to you that the Nazi extermination of the Jews never happened?" 22% said it was possible the Holocaust never happened. The Roper organization later admitted that the question was worded in a confusing manner. When the poll asked, "Does it seem possible to you that the Nazi extermination of the Jews never happened, or do you feel certain that it happened?" only 1%

said it was possible it never happened.[5] Use this example to explain the concept of response bias.

2.18. Refer to Exercise 2.12 about selecting 5 of 52,000 names on 400 pages of a directory.
- **(a)** Select five numbers to identify subjects for a systematic random sample of five names from the directory.
- **(b)** Is cluster sampling applicable? How could it be carried out, and what would be the advantages and disadvantages?

2.19. You plan to sample from the 5000 students at a college, to compare the proportions of men and women who believe that the legal age for alcohol should be changed to 18. Explain how you would proceed if you want a systematic random sample of 100 students.

2.20. You plan to sample from the 3500 undergraduate students enrolled at the University of Rochester, to compare the proportions of female and male students who would like to see the U.S. have a female President.
- **(a)** Suppose that you use random numbers to select students, but you stop selecting females as soon as you have 40, and you stop selecting males as soon as you have 40. Is the resulting sample a simple random sample? Why or why not?
- **(b)** What type of sample is the sample in (a)? What advantage might it have over a simple random sample?

2.21. Clusters versus strata:
- **(a)** With a cluster random sample, do you take a sample of (i) the clusters? (ii) the subjects within every cluster?
- **(b)** With a stratified random sample, do you take a sample of (i) the strata? (ii) the subjects within every stratum?
- **(c)** Summarize the main differences between cluster sampling and stratified sampling in terms of whether you sample the groups or sample from within the groups that form the clusters or strata.

Concepts and Applications

2.22. Refer to the *Student survey* data file introduced in Exercise 1.11 (page 8). For each variable in the data set, indicate whether it is:
- **(a)** Categorical or quantitative
- **(b)** Nominal, ordinal, or interval

2.23. Repeat the previous exercise for the data file created in Exercise 1.12 (page 9).

2.24. You are directing a study to determine the factors that relate to good academic performance at your school.
- **(a)** Describe how you might select a sample of 100 students for the study.
- **(b)** List some variables that you would measure. For each, provide the scale you would use to measure it, and indicate whether statistical analysis could treat it as (i) categorical or quantitative, (ii) nominal, ordinal, or interval, (iii) continuous or discrete.
- **(c)** Give an example of a research question that could be addressed using data on the variables you listed in (b).

2.25. With ***quota sampling*** a researcher stands at a street corner and conducts interviews until obtaining a quota representing the relative sizes of various groups in the population. For instance, the quota might be 50 factory workers, 100 housewives, 60 elderly people, 30 blacks, and so forth. Is this a probability or nonprobability sampling method? Explain, and discuss potential advantages or disadvantages of this method. (Professional pollsters such as Gallup used this method until 1948, when they incorrectly predicted that Dewey would defeat Truman in a landslide in the presidential election.)

2.26. When the Yankelovich polling organization asked,[6] "Should laws be passed to eliminate all possibilities of special interests giving huge sums of money to candidates?" 80% of the sample answered *yes*. When they posed the question, "Should laws be passed to prohibit interest groups from contributing to campaigns, or do groups have a right to contribute to the candidate they support?" only 40% said *yes*. Explain what this example illustrates, and use your answer to differentiate between sampling error and response bias in survey results.

2.27. In each of the following situations, evaluate whether the method of sample selection is appropriate for obtaining information about the population of interest. How would you improve the sample design?
- **(a)** A newspaper wants to determine whether its readers believe that government expenditures should be reduced by cutting benefits for the disabled. They provide an Internet address for readers to vote *yes* or *no*. Based on 1434 Internet votes, they report that 93% of the city's residents believe that benefits should be reduced.
- **(b)** A congresswoman reports that letters to her office are running 3 to 1 in opposition to

[5] *Newsweek*, July 25, 1994.

[6] *Source*: *A Mathematician Reads the Newspaper*, by J. A. Paulos, Basic Books, 1995, p. 15.

the passage of stricter gun control laws. She concludes that approximately 75% of her constituents oppose stricter gun control laws.

(c) An anthropology professor wanted to compare attitudes toward premarital sex of physical science majors and social science majors. She administered a questionnaire to her large class of Anthropology 437, Comparative Human Sexuality. She found no appreciable difference between her physical science and social science majors in their attitudes, so she concluded that the two student groups were about the same in their relative acceptance of premarital sex.

(d) A questionnaire was mailed to a simple random sample of 500 household addresses in a city. Ten were returned as bad addresses, 63 were returned completed, and the rest were not returned. The researcher analyzed the 63 cases and reported that they represent a "simple random sample of city households."

(e) A principal in a large high school is interested in student attitudes toward a proposed achievement test to determine whether a student should graduate. She lists all of the first-period classes, assigning a number to each. Then, using a random number table, she chooses a class at random and interviews every student in that class about the proposed test.

2.28. A content analysis of a daily newspaper studies the percentage of newspaper space devoted to news about entertainment. The sampling frame consists of the daily editions of the newspaper for the previous year. What potential problem might there be in using a systematic sample with skip number equal to 7 or a multiple of 7?

2.29. In a systematic random sample, every subject has the same chance of selection, but the sample is not a simple random sample. Explain why.

2.30. With a total sample of size 100, we want to compare Native Americans to other Americans on the percentage favoring legalized gambling. Why might it be useful to take a disproportional stratified random sample?

2.31. In a cluster random sample with equal-sized clusters, every subject has the same chance of selection. However, the sample is not a simple random sample. Explain why not.

2.32. Find an example of results of an Internet poll. Do you trust the results of the poll? If not, explain why not.

2.33. To sample residents of registered nursing homes in Yorkshire, UK, I construct a list of all nursing homes in the county, which I number from 1 to 110. Beginning randomly, I choose every tenth home on the list, ending up with 11 homes. I then obtain lists of residents from those 11 homes, and I select a simple random sample from each list. What kinds of sampling have I used?

For multiple-choice questions 2.34–2.37, select the best response.

2.34. A simple random sample of size n is one in which:

(a) Every nth member is selected from the population.

(b) Each possible sample of size n has the same chance of being selected.

(c) There must be exactly the same proportion of women in the sample as is in the population.

(d) You keep sampling until you have a fixed number of people having various characteristics (e.g., males, females).

(e) A particular minority group member of the population is less likely to be chosen than a particular majority group member.

(f) All of the above

(g) None of the above

2.35. If we use random numbers to take a simple random sample of 50 students from the 20,000 students at a university,

(a) It is impossible to get the random number 11111, because it is not a random sequence.

(b) If we get 20001 for the first random number, for the second random number that number is less likely to occur than the other possible five-digit random numbers.

(c) The draw 12345 is no more or less likely than the draw 11111.

(d) Since the sample is random, it is *impossible* that it will be non-representative, such as having only females in the sample.

2.36. Crosson (1994, p. 168) described an analysis of published medical studies involving treatments for heart attacks. In the studies having randomization and strong controls for bias, the new therapy provided improved treatment 9% of the time. In studies without randomization or other controls for bias, the new therapy provided improved treatment 58% of the time. Select the correct response(s).

(a) This result suggests it is better not to use randomization in medical studies, because it is harder to show that new ideas are beneficial.

(b) Many newspaper articles that suggest that a particular food, drug, or environmental agent is harmful or beneficial should be viewed skeptically, unless we learn more about the statistical design and analysis for the study.

(c) This result suggests that you should be skeptical about published results of medical studies that are not randomized, controlled studies.

(d) Controlling for biases, both suspected and unsuspected, is necessary in medical research but not in social research, because the social sciences deal in subjective rather than objective truth.

2.37. A recent GSS asked subjects if they supported legalizing abortion in each of seven different circumstances. The percentage who supported legalization varied between 45% (if the woman wants it for any reason) to 92% (if the woman's health is seriously endangered by the pregnancy). This indicates that

(a) Responses can depend greatly on the question wording.

(b) Surveys sample only a small part of the population and can never be trusted.

(c) The sample must not have been randomly selected.

(d) The sample must have had problems with bias resulting from subjects not telling the truth.

2.38. An interviewer stands at an entrance to a popular shopping mall and conducts interviews. True or false: Because we cannot predict who will be interviewed, the sample obtained is an example of a random sample. Explain.

2.39. In a recent Miss America beauty pageant, television viewers could cast their vote on whether to cancel the swimwear parade by phoning a number the network provided. About 1 million viewers called and registered their opinion, of whom 79% said they wanted to see the contestants dressed as bathing beauties. True or false: Since everyone had a chance to call, this was a simple random sample of all the viewers of this program. Explain.

2.40.** An interval scale for which ratios are valid is called a ***ratio scale. Such scales have a well-defined 0 point, so, for instance, one can regard the value 20 as twice the quantity of the value 10. Explain why annual income is measured on a ratio scale, but temperature (in Fahrenheit or Centigrade) is not. Is IQ, as a measure of intelligence, a ratio-scale variable?

*Exercises marked with an asterisk are of greater difficulty or introduce new and optional material.

This page intentionally left blank

CHAPTER 3

Descriptive Statistics

We've seen that statistical methods are *descriptive* or *inferential*. The purpose of descriptive statistics is to summarize data, to make it easier to assimilate the information. This chapter presents basic methods of descriptive statistics.

We first present tables and graphs that describe the data by showing the number of times various outcomes occurred. Quantitative variables also have two key features to describe numerically:

- The ***center*** of the data—a typical observation
- The ***variability*** of the data—the spread around the center

We'll learn how to describe quantitative data with statistics that summarize the center, statistics that summarize the variability, and finally with statistics that specify certain positions in the data set that summarize both center and variability.

3.1 DESCRIBING DATA WITH TABLES AND GRAPHS

Tables and graphs are useful for all types of data. We'll begin with categorical variables.

Relative Frequencies: Categorical Data

For categorical data, we list the categories and show the frequency (the number of observations) in each category. To make is easier to compare different categories, we also report proportions or percentages, also called ***relative frequencies***.

Relative Frequency

The ***relative frequency*** for a category is the ***proportion*** or ***percentage*** of the observations that fall in that category.

The *proportion* equals the number of observations in a category divided by the total number of observations. It is a number between 0 and 1 that expresses the share of the observations in that category. The *percentage* is the proportion multiplied by 100.

EXAMPLE 3.1 Household Structure in the U.S.

Table 3.1 lists the different types of households in the United States in 2005. Of 111.1 million households, for example, 24.1 million were a married couple with children. The proportion $24.1/111.1 = 0.22$ were a married couple with children.

TABLE 3.1: U.S. Household Structure, 2005

Type of Family	Number (millions)	Proportion	Percentage
Married couple with children	24.1	0.22	22
Married couple, no children	31.1	0.28	28
Single householder, no spouse	19.1	0.17	17
Living alone	30.1	0.27	27
Other households	6.7	0.06	6
Total	111.1	1.00	100

Source: U.S. Census Bureau, *2005 American Community Survey,* Tables B11001, C11003.

A percentage is the proportion multiplied by 100. That is, the decimal place is moved two positions to the right. For example, since 0.22 is the proportion of families that are married couples with children, the percentage is $100(0.22) = 22\%$. Table 3.1 shows the proportions and percentages for all the categories. ■

The sum of the proportions equals 1.00. The sum of the percentages equals 100. (In practice, the values may sum to a slightly different number, such as 99.9 or 100.1, because of rounding.)

It is sufficient in such a table to report the percentages (or proportions) and the total sample size, since each frequency equals the corresponding proportion multiplied by the total sample size. For instance, the frequency of married couples with children equals $0.22(111.1) = 24$ million. When presenting the percentages but not the frequencies, always also include the total sample size.

Frequency Distributions and Bar Graphs: Categorical Data

Table 3.1 lists the categories for household structure and the number of households of each type. Such a listing is called a *frequency distribution*.

Frequency Distribution

A ***frequency distribution*** is a listing of possible values for a variable, together with the number of observations at each value. A corresponding ***relative frequency distribution*** lists the possible values together with their proportions or percentages.

To construct a frequency distribution for a categorical variable, list the categories and count the number of observations in each.

To more easily get a feel for the data, it's helpful to look at a graph of the relative frequency distribution. A ***bar graph*** has a rectangular bar drawn over each category. The height of the bar shows the relative frequency in that category. Figure 3.1 is a bar graph for the data in Table 3.1. The bars are separated to emphasize that the variable is categorical rather than quantitative. Since household structure is a nominal variable, there is no particular natural order for the bars. The order of presentation for an ordinal variable is the natural ordering of the categories.

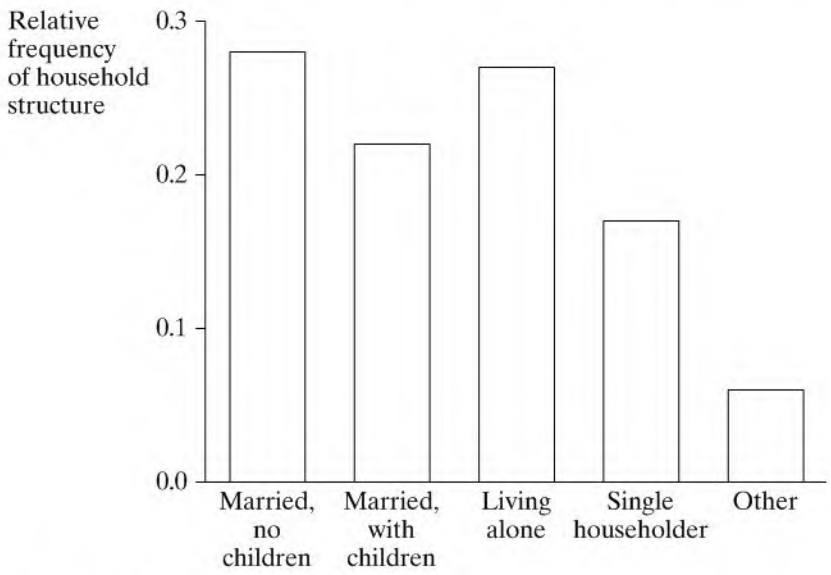


FIGURE 3.1: Relative Frequency of U.S. Household Structure Types, 2005

Another type of graph, the *pie chart*, is a circle having a "slice of the pie" for each category. The size of a slice represents the percentage of observations in the category. The bar graph is more precise than the pie chart for visual comparison of categories with similar relative frequencies.

Frequency Distributions: Quantitative Data

Frequency distributions and graphs also are useful for quantitative variables. The next example illustrates.

EXAMPLE 3.2 Statewide Violent Crime Rates

Table 3.2 lists all 50 states in the United States and their 2005 violent crime rates. This rate measures the number of violent crimes in that state in 2005 per 10,000 population. For instance, if a state had 12,000 violent crimes and a population size of 2,300,000, its violent crime rate was $(12{,}000/2{,}300{,}000) \times 10{,}000 = 52$. It is difficult to learn much by simply reading through the violent crime rates. Tables, graphs, and numerical measures help us more fully absorb the information in these data.

First, we can summarize the data with a frequency distribution. To do this, we divide the measurement scale for violent crime rate into a set of intervals and count the number of observations in each interval. Here, we use the intervals {0–11, 12–23, 24–35, 36–47, 48–59, 60–71, 72–83}. The values Table 3.2 reports were rounded, so for example the interval 12-23 represents values between 11.5 and 23.5. Counting the number of states with violent crime rates in each interval, we get the frequency distribution shown in Table 3.3. We see that considerable variability exists in the violent crime rates.

Table 3.3 also shows the relative frequencies, using proportions and percentages. For example, $3/50 = 0.06$ is the proportion for the interval 0–11, and $100(0.06) = 6$ is the percentage. As with any summary method, we lose some information as the cost of achieving some clarity. The frequency distribution does not identify which states have low or high violent crime rates, nor are the exact violent crime rates known. ■

The intervals of values in frequency distributions are usually of equal width. The width equals 12 in Table 3.3. The intervals should include all possible values of the

TABLE 3.2: List of States with Violent Crime Rates Measured as Number of Violent Crimes per 10,000 Population

Alabama	43	Louisiana	65	Ohio	33
Alaska	59	Maine	11	Oklahoma	51
Arizona	51	Maryland	70	Oregon	30
Arkansas	46	Massachusetts	47	Pennsylvania	40
California	58	Michigan	51	Rhode Island	29
Colorado	34	Minnesota	26	South Carolina	79
Connecticut	31	Mississippi	33	South Dakota	17
Delaware	66	Missouri	47	Tennessee	69
Florida	73	Montana	36	Texas	55
Georgia	45	Nebraska	29	Utah	25
Hawaii	27	Nevada	61	Vermont	11
Idaho	24	New Hampshire	15	Virginia	28
Illinois	56	New Jersey	37	Washington	35
Indiana	35	New Mexico	66	West Virginia	26
Iowa	27	New York	46	Wisconsin	22
Kansas	40	North Carolina	46	Wyoming	26
Kentucky	26	North Dakota	8		

TABLE 3.3: Frequency Distribution and Relative Frequency Distribution for Violent Crime Rates

Violent Crime Rate	Frequency	Relative Frequency	Percentage
0–11	3	0.06	6
12–23	3	0.06	6
24–35	18	0.36	36
36–47	11	0.22	22
48–59	7	0.14	14
60–71	6	0.12	12
72–83	2	0.04	4
Total	50	1.00	100.0

variable. In addition, any possible value must fit into one and only one interval; that is, they should be ***mutually exclusive***.

Histograms

A graph of a relative frequency distribution for a quantitative variable is called a ***histogram***. Each interval has a bar over it, with height representing the number of observations in that interval. Figure 3.2 is a histogram for the violent crime rates.

Choosing intervals for frequency distributions and histograms is primarily a matter of common sense. If too few intervals are used, too much information is lost. For example, Figure 3.3 is a histogram of violent crime rates using the intervals 0–29, 30–59, 60–89. This is too crude to be very informative. If too many intervals are used, they are so narrow that the information presented is difficult to digest, and the histogram may be irregular and the overall pattern of the results may be obscured. Ideally, two observations in the same interval should be similar in a practical sense. To summarize annual income, for example, if a difference of $5000 in income is not considered practically important, but a difference of $15,000 is notable, we might

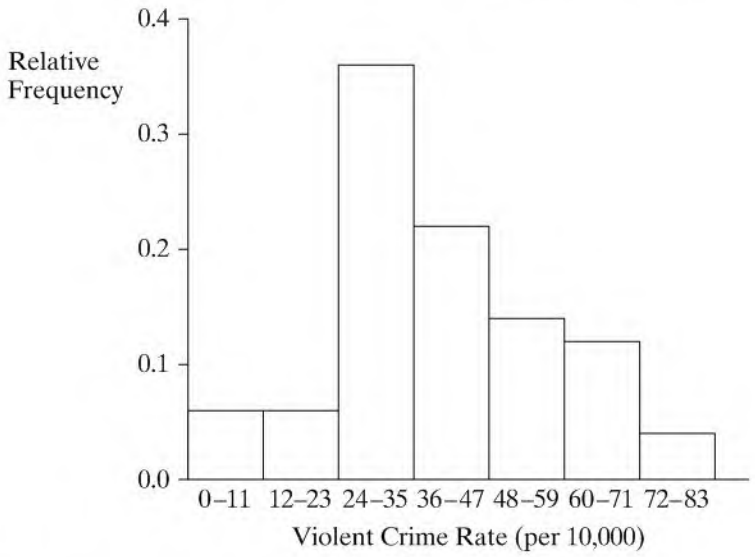


FIGURE 3.2: Histogram of Relative Frequencies for Statewide Violent Crime Rates

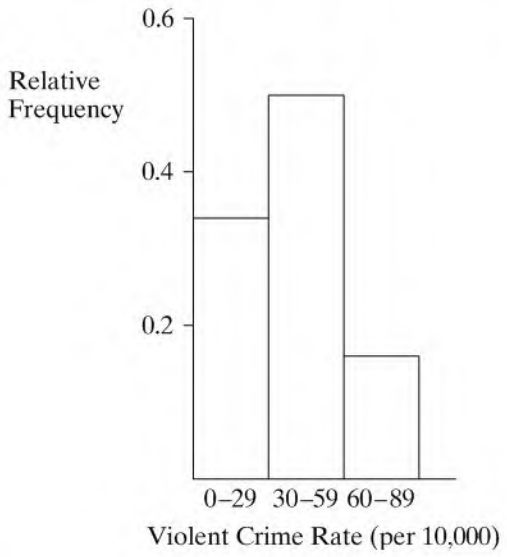


FIGURE 3.3: Histogram of Relative Frequencies for Violent Crime Rates, Using Too Few Intervals

choose intervals of width less than $15,000, such as $0–$9999, $10,000–$19,999, $20,000–$29,999, and so forth. Statistical software can automatically choose intervals for us and construct frequency distributions and histograms.

For a discrete variable with relatively few values, a histogram has a separate bar for each possible value. For a continuous variable or a discrete variable with many possible values, you need to divide the possible values into intervals, as we did with the violent crime rates.

Stem-and-Leaf Plots

Figure 3.4 shows an alternative graphical representation of the violent crime rate data. This figure, called a ***stem-and-leaf plot***, represents each observation by its leading digit(s) (the *stem*) and by its final digit (the *leaf*). Each stem is a number to the left of the vertical bar and a leaf is a number to the right of it. For instance, on the second line, the stem of 1 and the leaves of 1, 1, 5, and 7 represent the violent crime rates 11, 11, 15, 17. The plot arranges the leaves in order on each line, from smallest to largest.

Stem	Leaf											
0	8											
1	1	1	5	7								
2	2	4	5	6	6	6	6	7	7	8	9	9
3	0	1	3	3	4	5	5	6	7			
4	0	0	3	5	6	6	6	7	7			
5	1	1	1	5	6	8	9					
6	1	5	6	6	9							
7	0	3	9									

FIGURE 3.4: Stem-and-Leaf Plot for Violent Crime Rate Data in Table 3.2

A stem-and-leaf plot conveys similar information as a histogram. Turned on its side, it has the same shape as the histogram. In fact, since the stem-and-leaf plot shows each observation, it displays information that is lost with a histogram. From Figure 3.4, the largest violent crime rate was 79 and the smallest was 8 (shown as 08 with a stem of 0 and leaf of 8). It is not possible to determine these exact values from the histogram in Figure 3.2.

Stem-and-leaf plots are useful for quick portrayals of small data sets. As the sample size increases, you can accommodate the increase in leaves by splitting the stems. For instance, you can list each stem twice, putting leaves of 0 to 4 on one line and leaves of 5 to 9 on another. When a number has several digits, it is simplest for graphical portrayal to drop the last digit or two. For instance, for a stem-and-leaf plot of annual income in thousands of dollars, a value of \$27.1 thousand has a stem of 2 and a leaf of 7 and a value of \$106.4 thousand has a stem of 10 and leaf of 6.

Comparing Groups

Many studies compare different groups on some variable. Relative frequency distributions, histograms, and stem-and-leaf plots are useful for making comparisons.

EXAMPLE 3.3 Comparing Canadian and U.S. Murder Rates

Stem-and-leaf plots can provide visual comparisons of two small samples on a quantitative variable. For ease of comparison, the results are plotted "back to back." Each plot uses the same stem, with leaves for one sample to its left and leaves for the other sample to its right. To illustrate, Figure 3.5 shows back-to-back stem and leaf plots of recent murder rates (measured as the number of murders per 100,000 population) for the 50 states in the U.S. and for the provinces of Canada. From this figure, it is clear that the murder rates tended to be much lower in Canada, varying between 0.7 (Prince Edward Island) and 2.9 (Manitoba) whereas those in the U.S. varied between 1.6 (Maine) and 20.3 (Louisiana). ■

Population Distribution and Sample Data Distribution

Frequency distributions and histograms apply both to a population and to samples from that population. The first type is called the ***population distribution***, and the second type is called a ***sample data distribution***. In a sense, the sample data distribution is a blurry photo of the population distribution. As the sample size increases, the sample proportion in any interval gets closer to the true population proportion. Thus, the sample data distribution looks more like the population distribution.

Canada						Stem	United States									
					7	0										
			3	2	1	1	6	7								
9	7	6	3	2	0	2	0	3	9							
						3	0	1	4	4	4	6	8	9	9	9
						4	4	6								
						5	0	2	3	8						
						6	0	3	4	6	8	9				
						7	5									
						8	0	3	4	6	9					
						9	0	8								
						10	2	2	3	4						
						11	3	3	4	4	6	9				
						12	7									
						13	1	3	5							
						14										
						15										
						16										
						17										
						18										
						19										
						20	3									

FIGURE 3.5: Back-to-Back Stem-and-Leaf Plots of Murder Rates from U.S. and Canada. Both share the same stems, with Canada leafs to the left and U.S. leafs to the right.

For a continuous variable, imagine the sample size increasing indefinitely, with the number of intervals simultaneously increasing, so their width narrows. Then, the shape of the sample histogram gradually approaches a smooth curve. This text uses such curves to represent population distributions. Figure 3.6 shows two sample histograms, one based on a sample of size 100 and the second based on a sample of size 500, and also a smooth curve representing the population distribution. Even if a variable is discrete, a smooth curve often approximates well the population distribution, especially when the number of possible values of the variable is large.

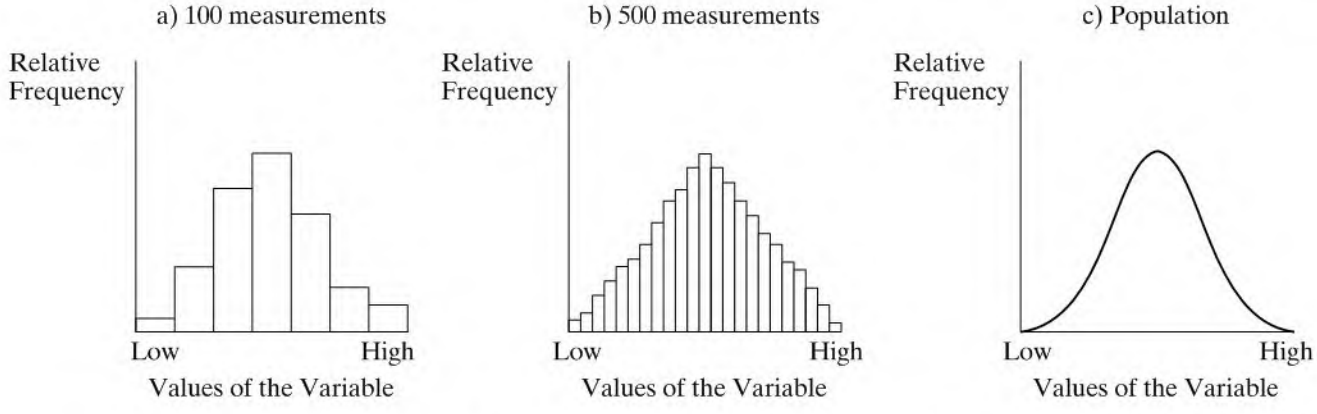


FIGURE 3.6: Histograms for a Continuous Variable. We use smooth curves to represent population distributions for continuous variables.

The Shape of a Distribution

One way to summarize a sample or a population distribution is to describe its shape. A group for which the distribution is bell-shaped is fundamentally different from

a group for which the distribution is U-shaped, for example. See Figure 3.7. In the U-shaped distribution, the highest points (representing the largest frequencies) are at the lowest and highest scores, whereas in the bell-shaped distribution, the highest point is near the middle value. A U-shaped distribution indicates a polarization on the variable between two sets of subjects. A bell-shaped distribution indicates that most subjects tend to fall near a central value.

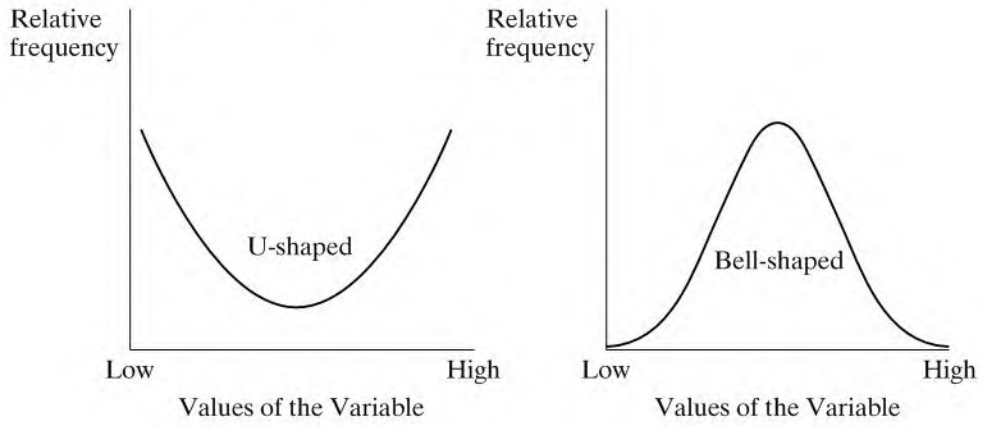


FIGURE 3.7: U-Shaped and Bell-Shaped Frequency Distributions

The distributions in Figure 3.7 are ***symmetric***: The side of the distribution below a central value is a mirror image of the side above that central value. Most distributions encountered in the social sciences are not symmetric. Figure 3.8 illustrates. The parts of the curve for the lowest values and the highest values are called the ***tails*** of the distribution. Often, as in Figure 3.8, one tail is much longer than the other. A distribution is said to be ***skewed to the right*** or ***skewed to the left***, according to which tail is longer.

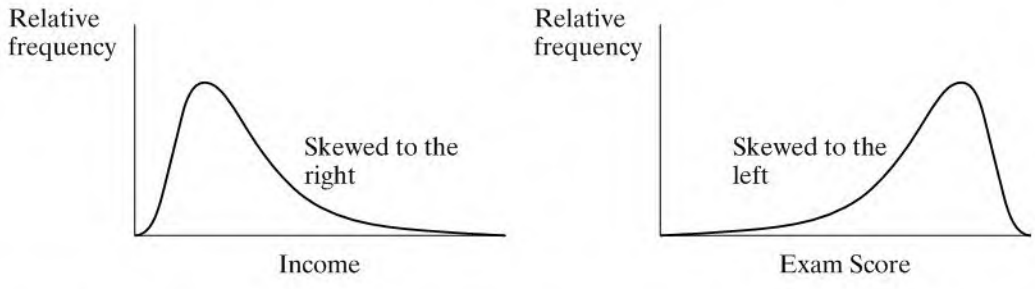


FIGURE 3.8: Skewed Frequency Distributions. The longer tail indicates the direction of skew.

To compare frequency distributions or histograms for two groups, you can give verbal descriptions using characteristics such as skew. It is also helpful to make numerical comparisons such as, "On the average, the murder rate for U.S. states is 5.4 higher than the murder rate for Canadian provinces." We now turn our attention to numerical descriptive statistics.

3.2 DESCRIBING THE CENTER OF THE DATA

This section presents statistics that describe the center of a frequency distribution for a quantitative variable. The statistics show what a *typical* observation is like.

The Mean

The best known and most commonly used measure of the center is the ***mean***.

Mean

The ***mean*** is the sum of the observations divided by the number of observations.

The mean is often called the ***average***.

EXAMPLE 3.4 Female Economic Activity in Europe

Table 3.4 shows an index of female economic activity for the countries of South America and of Eastern Europe in 2003. The number specifies female employment as a percentage of male employment. In Argentina, for instance, the number of females in the work force was 48% of the number of males in the work force. (The value was 83 in the United States and in Canada.)

TABLE 3.4: Female Economic Activity in South America and Eastern Europe; Female Employment as a Percentage of Male Employment

South America		Eastern Europe	
Country	Activity	Country	Activity
Argentina	48	Czech republic	83
Bolivia	58	Estonia	82
Brazil	52	Hungary	72
Chile	50	Latvia	80
Colombia	62	Lithuania	80
Ecuador	40	Poland	81
Guyana	51	Slovakia	84
Paraguay	44	Slovenia	81
Peru	45		
Uruguay	68		
Venezuela	55		

Source: Human Development Report 2005, United Nations Development Programme.

For the eight observations for Eastern Europe, the sum equals

$$83 + 82 + 72 + 80 + 80 + 81 + 84 + 81 = 643.$$

The mean female economic activity equals $643/8 = 80.4$. By comparison, you can check that the mean for the 11 South American countries equals $573/11 = 52.1$. Female economic activity tends to be considerably lower in South America than in Eastern Europe. ■

We use the following notation for the mean in formulas for it and for statistics that use the mean.

Notation for Observations and Sample Mean

The sample size is symbolized by n. For a variable denoted by y, its observations are denoted by $y_1, y_2, \ldots, y_n$. The sample mean is denoted by $\bar{y}$.

The symbol $\bar{y}$ for the sample mean is read as "y-bar." Throughout the text, letters near the end of the alphabet denote variables. The n sample observations on a variable y are denoted by y_1 for the first observation, y_2 the second, and so forth. For example, for female economic activity in Eastern Europe, $n = 8$ and the observations are $y_1 = 83$, $y_2 = 82, \ldots, y_8 = 81$. A bar over a letter represents the sample mean for that variable. For instance, $\bar{x}$ represents the sample mean for a variable denoted by x.

The definition of the sample mean says that

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}.$$

The symbol Σ (uppercase Greek letter sigma) represents the process of summing. For instance, Σy_i represents the sum $y_1 + y_2 + \cdots + y_n$. This symbol stands for the sum of the y-values, where the index i represents a typical value in the range 1 to n. To illustrate, for the Eastern European data,

$$\sum y_i = y_1 + y_2 + \cdots + y_8 = 83 + 82 + \cdots + 81 = 643.$$

The symbol is sometimes even further abbreviated as Σy. Using this summation symbol, we have the shortened expression for the sample mean of n observations,

$$\bar{y} = \frac{\Sigma y_i}{n}.$$

Properties of the Mean

Here are some properties of the mean:

- The formula for the mean uses numerical values for the observations. So the mean is appropriate only for quantitative variables. It is not sensible to compute the mean for observations on a nominal scale. For instance, for religion measured with categories such as (Protestant, Catholic, Jewish, Other), the mean religion does not make sense, even though these levels may sometimes be coded by numbers for convenience. Similarly, we cannot find the mean of observations on an ordinal rating such as excellent, good, fair, and poor, unless we assign numbers such as 4, 3, 2, 1 to the ordered levels, treating it as quantitative.
- The mean can be highly influenced by an observation that falls well above or well below the bulk of the data, called an ***outlier***.

EXAMPLE 3.5 Effect of Outlier on Mean Income

The owner of Leonardo's Pizza reports that the mean annual income of employees in the business is \$40,900. In fact, the annual incomes of the seven employees are \$11,200, \$11,400, \$11,700, \$12,200, \$12,300, \$12,500, and \$215,000. The \$215,000 income is the salary of the owner's son, who happens to be an employee. The value \$215,000 is an outlier. The mean computed for the other six observations alone equals \$11,883, quite different from the mean of \$40,900 including the outlier. ■

This example shows that the mean is not always typical of the observations in the sample. This commonly happens with small samples when at least one observation is

much larger or much smaller than the others, such as in highly skewed distributions.

- The mean is pulled in the direction of the longer tail of a skewed distribution, relative to most of the data.

  In Example 3.5, the large observation \$215,000 results in an extreme skewness to the right of the income distribution. This skewness pulls the mean above six of the seven observations. In general, the more highly skewed the distribution, the less typical the mean is of the data.

- The mean is the point of balance on the number line when an equal weight is at each observation point.

  For example, Figure 3.9 shows that if an equal weight is placed at each Eastern European observation on female economic activity from Example 3.4, then the line balances by placing a fulcrum at the point 80.4. The mean is the *center of gravity* (balance point) of the observations. This means that the sum of the distances to the mean from the observations *above* the mean equals the sum of the distances to the mean from the observations *below* the mean.

FIGURE 3.9: The Mean as the Center of Gravity, for Eastern Europe Data from Example 3.4. The line balances with a fulcrum at 80.4.

- Denote the sample means for two sets of data with sample sizes n_1 and n_2 by $\bar{y}_1$ and $\bar{y}_2$. The overall sample mean for the combined set of $(n_1 + n_2)$ observations is the ***weighted average***

$$\bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2}.$$

  The numerator $n_1\bar{y}_1 + n_2\bar{y}_2$ is the sum of all the observations, since $n\bar{y} = \sum y$ for each set of observations. The denominator is the total sample size.

To illustrate, for the female economic activity data in Table 3.4, the South American observations have $n_1 = 11$ and $\bar{y}_1 = 52.1$, and the Eastern European observations have $n_2 = 8$ and $\bar{y}_2 = 80.4$. The overall mean economic activity for the 19 nations equals

$$\bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2} = \frac{11(52.1) + 8(80.4)}{11 + 8} = \frac{(573 + 643)}{19} = \frac{1216}{19} = 64.$$

The weighted average of 64 is closer to 52.1, the value for South America, than to 80.4, the value for Eastern Europe. This happens because more observations come from South America than Eastern Europe.

The Median

The mean is a simple measure of the center. But other measures are also informative and sometimes more appropriate. Most important is the *median*. It splits the sample into two parts with equal numbers of observations, when they are ordered from lowest to highest.

Median

The ***median*** is the observation that falls in the middle of the ordered sample. When the sample size n is odd, a single observation occurs in the middle. When the sample size is even, two middle observations occur, and the median is the midpoint between the two.

To illustrate, the ordered income observations for the seven employees in Example 3.5 are

$$\$11{,}200, \$11{,}400, \$11{,}700, \$12{,}200, \$12{,}300, \$12{,}500, \$215{,}000.$$

The median is the middle observation, \$12,200. This is a more typical value for this sample than the sample mean of \$40,900. When a distribution is highly skewed, the median describes a typical value better than the mean.

In Table 3.4, the ordered economic activity values for the Eastern European nations are

$$72, 80, 80, 81, 81, 82, 83, 84.$$

Since $n = 8$ is even, the median is the midpoint between the two middle values, 81 and 81, which is $(81 + 81)/2 = 81$. This is close to the sample mean of 80.4, because this data set has no outliers.

The middle observation has index $(n + 1)/2$. That is, the median is the value of observation $(n + 1)/2$ in the ordered sample. When $n = 7$, $(n + 1)/2 = (7 + 1)/2 = 4$, so the median is the fourth smallest, or equivalently fourth largest, observation. When n is even, $(n + 1)/2$ falls halfway between two numbers, and the median is the midpoint of the observations with those indices. For example, when $n = 8, (n + 1)/2 = 4.5$, so the median is the midpoint between the 4th and 5th smallest observations.

EXAMPLE 3.6 Median for Grouped or Ordinal Data

Table 3.5 summarizes the distribution of the highest degree completed in the U.S. population of age 25 years and over, as estimated from the 2005 American Community Survey taken by the U.S. Bureau of the Census. The possible responses form an ordinal scale. The population size was $n = 189$ (in millions). The median score is the $(n + 1)/2 = (189 + 1)/2 = 95$th lowest. Now 30 responses fall in the first category, $(30 + 56) = 86$ in the first two, $(30 + 56 + 38) = 124$ in the first three, and so forth. The 87th to 124th lowest scores fall in category 3, which therefore contains the 95th lowest, which is the median. The median response is "Some college, no degree." Equivalently, from the percentages in the last column of the table, $(15.9\% + 29.6\%) = 45.5\%$ fall in the first two categories and $(15.9\% + 29.6\% + 20.1\%) = 65.6\%$ fall in the first three, so the 50% point falls in the third category. ■

TABLE 3.5: Highest Degree Completed, for a Sample of Americans

Highest Degree	Frequency (millions)	Percentage
Not a high school graduate	30	15.9%
High school only	56	29.6%
Some college, no degree	38	20.1%
Associate's degree	14	7.4%
Bachelor's degree	32	16.9%
Master's degree	13	6.9%
Doctorate or professional	6	3.2%

Properties of the Median

- The median, like the mean, is appropriate for quantitative variables. Since it requires only ordered observations to compute it, it is also valid for ordinal-scale data, as the previous example showed. It is not appropriate for nominal-scale data, since the observations cannot be ordered.
- For symmetric distributions, such as in Figure 3.7, the median and the mean are identical. To illustrate, the sample of observations 4, 5, 7, 9, 10 is symmetric about 7; 5 and 9 fall equally distant from it in opposite directions, as do 4 and 10. Thus, 7 is both the median and the mean.
- For skewed distributions, the mean lies toward the direction of skew (the longer tail) relative to the median. See Figure 3.10.

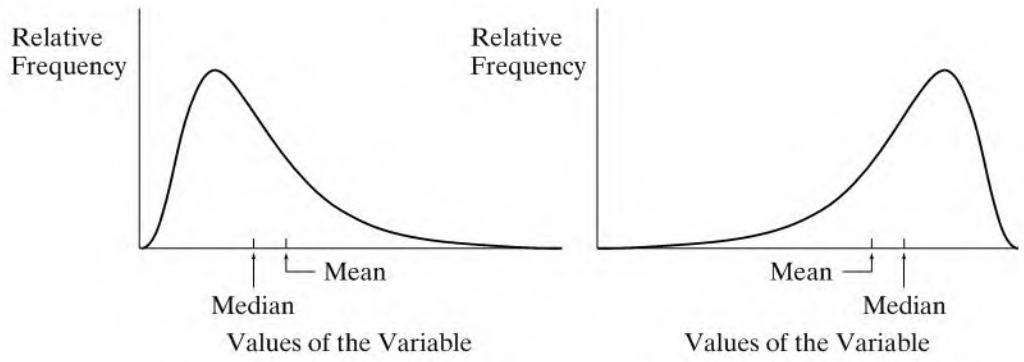


FIGURE 3.10: The Mean and the Median for Skewed Distributions. The mean is pulled in the direction of the longer tail.

For example, consider the violent crime rates of Table 3.2. The median is 36.5. The mean is $\bar{y} = 40.2$, somewhat larger than the median. Figure 3.2 showed that the violent crime rate values are skewed to the right. The mean is larger than the median for distributions that are skewed to the right. Income distributions tend to be skewed to the right. For example, household income in the United States in 2005 had a mean of about $61,000 and a median of about $44,000 (U.S. Bureau of the Census).

The distribution of grades on an exam tends to be skewed to the left when some students perform considerably poorer than the others. In this case, the mean is less than the median. For example, suppose that an exam scored on a scale of 0 to 100 has a median of 88 and a mean of 76. Then most students performed quite well (half being over 88), but apparently some scores were very much lower in order to bring the mean down to 76.

- The median is insensitive to the distances of the observations from the middle, since it uses only the ordinal characteristics of the data. For example, the following four sets of observations all have medians of 10:

Set 1:	8,	9,	10,	11,	12
Set 2:	8,	9,	10,	11,	100
Set 3:	0,	9,	10,	10,	10
Set 4:	8,	9,	10,	100,	100

- The median is not affected by outliers. For instance, the incomes of the seven employees in Example 3.5 have a median of $12,200 whether the largest observation is $20,000, $215,000, or $2,000,000.

Median Compared to Mean

The median is usually more appropriate than the mean when the distribution is highly skewed, as we observed with the Leonardo's Pizza employee incomes. The mean can be greatly affected by outliers, whereas the median is not.

For the mean we need quantitative (interval-scale) data. The median also applies for ordinal scales (see Example 3.6). To use the mean for ordinal data, we must assign scores to the categories. In Table 3.5, if we assign scores 10, 12, 13, 14, 16, 18, 20 to the categories of highest degree, representing approximate number of years of education, we get a sample mean of 13.4.

The median has its own disadvantages. For discrete data that take relatively few values, quite different patterns of data can have the same median. For instance, Table 3.6, from a GSS, summarizes the 365 female responses to the question, "How many sex partners have you had in the last 12 months?" Only six distinct responses occur, and 63.8% of those are 1. The median response is 1. To find the sample mean, to sum the 365 observations we multiply each possible value by the frequency of its occurrence, and then add. That is,

$$\sum y_i = 102(0) + 233(1) + 18(2) + 9(3) + 2(4) + 1(5) = 309.$$

The sample mean response is

$$\bar{y} = \frac{\Sigma y_i}{n} = \frac{309}{365} = 0.85.$$

If the distribution of the 365 observations among these categories were (0, 233, 18, 9, 2, 103) (i.e., we shift 102 responses from 0 to 5), then the median would still be 1, but the mean would shift to 2.2. The mean uses the numerical values of the observations, not just their ordering.

TABLE 3.6: Number of Sex Partners Last Year, for Female Respondents in GSS

Response	Frequency	Percentage
0	102	27.9
1	233	63.8
2	18	4.9
3	9	2.5
4	2	0.5
5	1	0.3

The most extreme form of this problem occurs for ***binary data***, which can take only two values, such as 0 and 1. The median equals the more common outcome, but gives no information about the relative number of observations at the two levels. For instance, consider a sample of size 5 for the variable, number of times married. The observations (1, 1, 1, 1, 1) and the observations (0, 0, 1, 1, 1) both have a median of 1. The mean is 1 for (1, 1, 1, 1, 1) and 3/5 for (0, 0, 1, 1, 1). *When observations take values of only 0 or 1, the mean equals the proportion of observations that equal 1.* Generally, for highly discrete data, the mean is more informative than the median.

In summary,

- If a distribution is highly skewed, the median is usually preferred because it better represents what is typical.

- If the distribution is close to symmetric or only mildly skewed or if it is discrete with few distinct values, the mean is usually preferred, because it uses the numerical values of all the observations.

The Mode

Another measure, the *mode*, indicates the most common outcome.

Mode

The ***mode*** is the value that occurs most frequently.

The mode is most commonly used with highly discrete variables, such as with categorical data. In Table 3.5, on the highest degree completed, for instance, the mode is "High school only," since the frequency for that category is higher than the frequency for any other rating. In Table 3.6, on the number of sex partners in the last year, the mode is 1.

Properties of the Mode

- The mode is appropriate for all types of data. For example, we might measure the mode for religion in Australia (nominal scale), for the rating given a teacher (ordinal scale), or for the number of years of education completed by Hispanic Americans (interval scale).
- A frequency distribution is called ***bimodal*** if two distinct mounds occur in the distribution. Bimodal distributions often occur with attitudinal variables when populations are polarized, with responses tending to be strongly in one direction or another. For instance, Figure 3.11 shows the relative frequency distribution of responses in a General Social Survey to the question, "Do you personally think it is wrong or not wrong for a woman to have an abortion if the family has a very low income and cannot afford any more children?" The relative frequencies in the two extreme categories are higher than those in the middle categories.

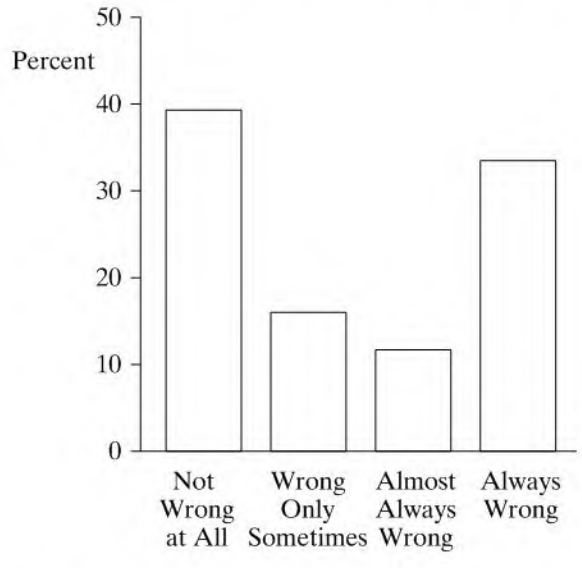


FIGURE 3.11: Bimodal Distribution for Opinion about Whether Abortion Is Wrong

- The mean, median, and mode are identical for a unimodal, symmetric distribution, such as a bell-shaped distribution.

The mean, median, and mode are complementary measures. They describe different aspects of the data. In any particular example, some or all their values may be useful. Be on the lookout for misleading statistical analyses, such as using one statistic when another would be more informative. People who present statistical conclusions often choose the statistic giving the impression they wish to convey. Recall Example 3.5 (p. 40) on Leonardo's Pizza employees, with the extreme outlying income observation. Be wary of the mean when the distribution may be highly skewed.

3.3 DESCRIBING VARIABILITY OF THE DATA

A measure of center alone is not adequate for numerically describing data for a quantitative variable. It describes a typical value, but not the spread of the data about that typical value. The two distributions in Figure 3.12 illustrate. The citizens of nation A and the citizens of nation B have the same mean annual income ($25,000). The distributions of those incomes differ fundamentally, however, nation B being much less variable. An income of $30,000 is extremely large for nation B, but not especially large for nation A. This section introduces statistics that describe the variability of a data set.

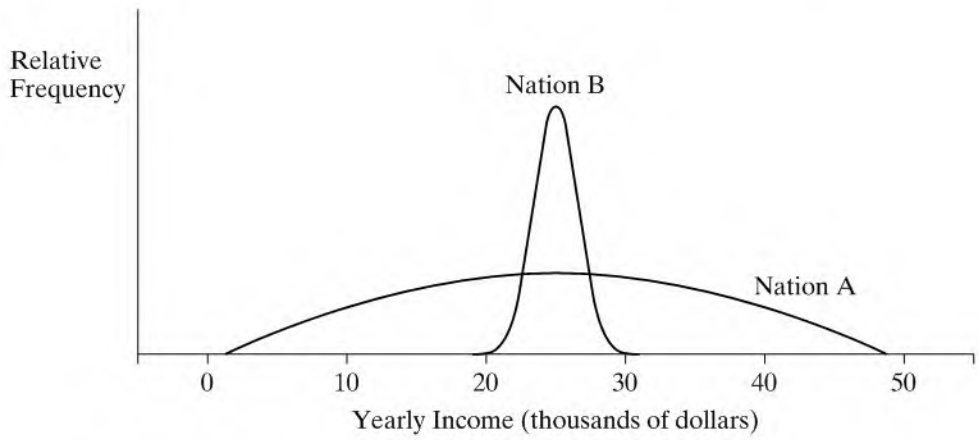


FIGURE 3.12: Distributions with the Same Mean but Different Variability

The Range

The difference between the largest and smallest observations is the simplest way to describe variability.

Range

The ***range*** is the difference between the largest and smallest observations.

For nation A, from Figure 3.12, the range of income values is about $50,000 − 0 = $50,000. For nation B, the range is about $30,000 − $20,000 = $10,000. Nation A has greater variability of incomes.

The range is not, however, sensitive to other characteristics of data variability. The three distributions in Figure 3.13 all have the same mean ($25,000) and range ($50,000), but they differ in variability about the center. In terms of distances of observations from the mean, nation A has the most variability, and nation B the least. The incomes in nation A tend to be farthest from the mean, and the incomes in nation B tend to be closest.

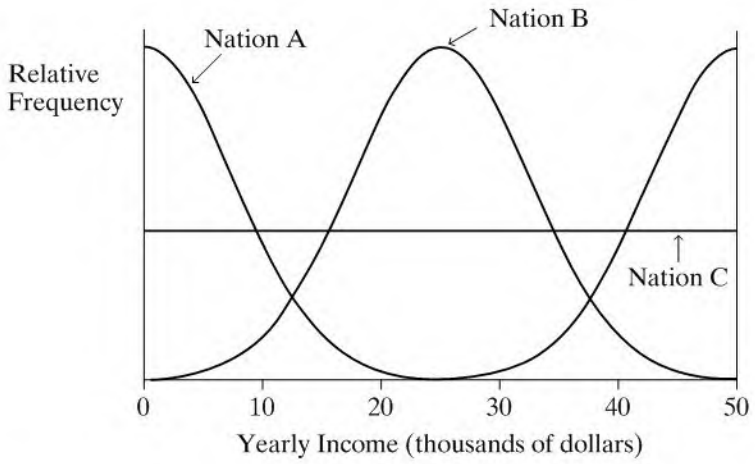


FIGURE 3.13: Distributions with the Same Mean and Range, but Different Variability about the Mean

Standard Deviation

Other measures of variability are based on the deviations of the data from a measure of center such as their mean.

Deviation

The ***deviation*** of an observation y_i from the sample mean y is $(y_i - y)$, the difference between them.

Each observation has a deviation. The deviation is *positive* when the observation falls *above* the mean. The deviation is *negative* when the observation falls *below* the mean. The interpretation of $\bar{y}$ as the center of gravity of the data implies that the sum of the positive deviations equals the negative of the sum of negative deviations. Thus, the sum of all the deviations about the mean, $\Sigma(y_i - \bar{y})$, equals 0. Because of this, measures of variability use either the absolute values or the squares of the deviations. The most popular measure uses the squares.

Standard Deviation

The ***standard deviation*** s of n observations is

$$s = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n - 1}} = \sqrt{\frac{\text{sum of squared deviations}}{\text{sample size} - 1}}.$$

This is the positive square root of the ***variance*** s^2, which is

$$s^2 = \frac{\Sigma(y_i - \bar{y})^2}{n - 1} = \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n - 1}.$$

The *variance* is approximately an average of the squared deviations. The units of measurement are the squares of those for the original data, since it uses squared deviations. This makes the variance difficult to interpret. It is why we use instead its square root, the *standard deviation*.

The expression $\Sigma(y_i - \bar{y})^2$ in these formulas is called a ***sum of squares***. It represents squaring each deviation and then adding those squares. It is incorrect to first add the deviations and then square that sum; this gives a value of 0. The larger the deviations, the larger the sum of squares and the larger s tends to be.

Although its formula looks complicated, the most basic interpretation of the standard deviation s is quite simple: s is a sort of *typical distance* of an observation from the mean. So the larger the standard deviation s, the greater the spread of the data.

EXAMPLE 3.7 Comparing Variability of Quiz Scores

Each of the following sets of quiz scores for two small samples of students has a mean of 5 and a range of 10:

Sample 1: 0, 4, 4, 5, 7, 10
Sample 2: 0, 0, 1, 9, 10, 10.

By inspection, sample 1 shows less variability about the mean than sample 2. Most scores in sample 1 are near the mean of 5, whereas all the scores in sample 2 are quite far from 5.

For sample 1,

$$\Sigma(y_i - \bar{y})^2 = (0 - 5)^2 + (4 - 5)^2 + (4 - 5)^2 + (5 - 5)^2 + (7 - 5)^2 + (10 - 5)^2 = 56,$$

so the variance equals

$$s^2 = \frac{\Sigma(y_i - \bar{y})^2}{n - 1} = \frac{56}{6 - 1} = \frac{56}{5} = 11.2.$$

The standard deviation for sample 1 equals $s = \sqrt{11.2} = 3.3$. For sample 2, you can verify that $s^2 = 26.4$ and $s = \sqrt{26.4} = 5.1$. Since $3.3 < 5.1$, the standard deviations tell us that sample 1 is less variable than sample 2. ■

Statistical software and many hand calculators can find the standard deviation. You should do the calculation yourself for a few small data sets to get a feel for what this measure represents. The answer you get may differ slightly from the value reported by software, depending on how much you round off in performing the calculation.

Properties of the Standard Deviation

- $s \geq 0$.
- $s = 0$ only when all observations have the same value. For instance, if the ages for a sample of five students are 19, 19, 19, 19, 19, then the sample mean equals 19, each of the five deviations equals 0, and $s = 0$. This is the minimum possible variability.
- The greater the variability about the mean, the larger is the value of s. For example, Figure 3.5 shows that murder rates are much more variable among U.S. states than among Canadian provinces. In fact, the standard deviations are $s = 4.0$ for the United States and $s = 0.8$ for Canada.
- The reason for using $(n - 1)$, rather than n, in the denominator of s (and s^2) is a technical one regarding inference about population parameters, discussed in Chapter 5. When we have data for an entire population, we replace $(n - 1)$ by the actual population size; the population variance is then precisely the mean of the squared deviations. In that case, the standard deviation can be no larger than half the range.
- If the data are rescaled, the standard deviation is also rescaled. For instance, if we change annual incomes from dollars (such as 34,000) to thousands of dollars (such as 34.0), the standard deviation also changes by a factor of 1000 (such as from 11,800 to 11.8).

Interpreting the Magnitude of *s*

A distribution with $s = 5.1$ has greater variability than one with $s = 3.3$, but how do we interpret *how large* $s = 5.1$ is? We've seen that a rough answer is that s is a typical distance of an observation from the mean. To illustrate, suppose the first exam in your course, graded on a scale of 0 to 100, has a sample mean of 77. A value of $s = 0$ in unlikely, since every student must then score 77. A value such as $s = 50$ seems implausibly large for a typical distance from the mean. Values of s such as 8 or 12 seem much more realistic.

More precise ways to interpret s require further knowledge of the shape of the frequency distribution. The following rule provides an interpretation for many data sets.

Empirical Rule

If the histogram of the data is approximately bell shaped, then

1. About 68% of the observations fall between $\bar{y} - s$ and $\bar{y} + s$.
2. About 95% of the observations fall between $\bar{y} - 2s$ and $\bar{y} + 2s$.
3. All or nearly all observations fall between $\bar{y} - 3s$ and $\bar{y} + 3s$.

The rule is called the Empirical Rule because many distributions seen in practice (that is, *empirically*) are approximately bell shaped. Figure 3.14 is a graphical portrayal of the rule.

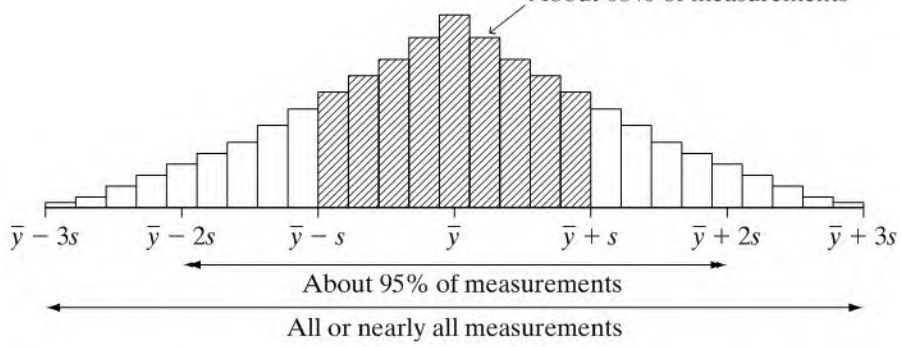


FIGURE 3.14: Empirical Rule: Interpretation of the Standard Deviation for a Bell-Shaped Distribution

EXAMPLE 3.8 Describing a Distribution of SAT Scores

The Scholastic Aptitude Test (SAT, see www.collegeboard.com) has three portions: Critical Reading, Mathematics, and Writing. For each portion, the distribution of scores is approximately bell shaped. Each portion has mean about 500 and standard deviation about 100. Figure 3.15 portrays this. By the Empirical Rule, for each portion, about 68% of the scores fall between 400 and 600, because 400 and 600 are the numbers that are *one* standard deviation below and above the mean of 500. About 95% of the scores fall between 300 and 700, the numbers that are *two* standard deviations from the mean. The remaining 5% fall either below 300 or above 700. The distribution is roughly symmetric about 500, so about 2.5% of the scores fall above 700 and about 2.5% fall below 300. ■

The Empirical Rule applies only to distributions that are approximately bell-shaped. For other shapes, the percentage falling within two standard deviations of the mean need not be near 95%. It could be as low as 75% or as high as 100%. The

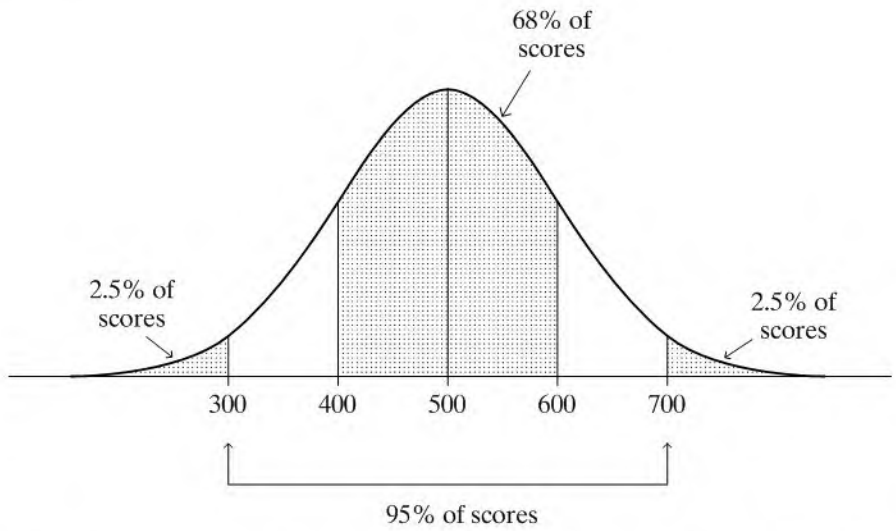


FIGURE 3.15: A Bell-Shaped Distribution of Scores for a Portion of the SAT, with Mean 500 and Standard Deviation 100

Empirical Rule may not work well if the distribution is highly skewed or if it is highly discrete, with the variable taking few values. The exact percentages depend on the form of the distribution, as the next example demonstrates.

EXAMPLE 3.9 Familiarity with AIDS Victims

A GSS asked, "How many people have you known personally, either living or dead, who came down with AIDS?" Table 3.7 shows part of a computer printout for summarizing the 1598 responses on this variable. It indicates that 76% of the responses were 0.

TABLE 3.7: Frequency Distribution of the Number of People Known Personally with AIDS

AIDS	Frequency	Percent
0	1214	76.0
1	204	12.8
2	85	5.3
3	49	3.1
4	19	1.2
5	13	0.8
6	5	0.3
7	8	0.5
8	1	0.1

N	1598
Mean	0.47
Std Dev	1.09

The mean and standard deviation are $\bar{y} = 0.47$ and $s = 1.09$. The values 0 and 1 both fall within one standard deviation of the mean. Now 88.8% of the distribution falls at these two points, or within $\bar{y} \pm s$. This is considerably larger than the 68% that the Empirical Rule states. The Empirical Rule does not apply to this distribution,

because it is not even approximately bell shaped. Instead, it is highly skewed to the right, as you can check by sketching a histogram for Table 3.7. The smallest value in the distribution (0) is less than one standard deviation below the mean; the largest value in the distribution (8) is nearly seven standard deviations above the mean. ■

Whenever the smallest or largest observation is less than a standard deviation from the mean, this is evidence of severe skew. For instance, a recent statistics exam having scale from 0 to 100 had $\bar{y} = 86$ and $s = 15$. The upper bound of 100 was less than one standard deviation above the mean. The distribution was highly skewed to the left.

The standard deviation, like the mean, can be greatly affected by an outlier, especially for small data sets. For instance, the murder rates shown in Figure 3.5 for the 50 U.S. states have $\bar{y} = 7.3$ and $s = 4.0$. The distribution is somewhat irregular, but 68% of the states have murder rates within one standard deviation of the mean and 98% within two standard deviations. Now suppose we include the murder rate for the District of Columbia, which equaled 78.5, in the data set. Then $\bar{y} = 8.7$ and $s = 10.7$. The standard deviation more than doubles. Now 96.1% of the murder rates (all except D.C. and Louisiana) fall within one standard deviation of the mean.

3.4 MEASURES OF POSITION

Another way to describe a distribution is with a measure of ***position***. This tells us the point at which a given percentage of the data fall below (or above) that point. As special cases, some measures of position describe center and some describe variability.

Quartiles and Other Percentiles

The range uses two measures of position, the maximum value and the minimum value. The median is a measure of position, with half the data falling below it and half above it. The median is a special case of a set of measures of position called *percentiles*.

Percentile

The ***pth percentile*** is the point such that p% of the observations fall below or at that point and $(100 - p)$% fall above it.

Substituting $p = 50$ in this definition gives the 50th percentile. This is the median. The median is larger than 50% of the observations and smaller than the other $(100 - 50) = 50$%. Two other commonly used percentiles are the *lower quartile* and the *upper quartile*.

Lower and Upper Quartiles

The 25th percentile is called the ***lower quartile***. The 75th percentile is called the ***upper quartile***. One quarter of the data fall below the lower quartile. One quarter fall above the upper quartile.

The quartiles result from $p = 25$ and $p = 75$ in the percentile definition. The lower quartile is the median for the observations that fall below the median, that is, for the bottom half of the data. The upper quartile is the median for the observations that fall above the median, that is, for the upper half of the data. The quartiles together with the median split the distribution into four parts, each containing one-fourth of the observations. See Figure 3.16.

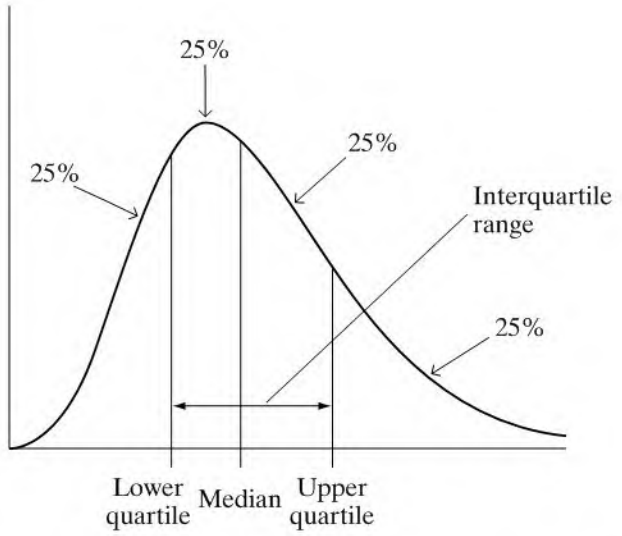


FIGURE 3.16: The Quartiles and the Interquartile Range

For the violent crime rates in Table 3.2, the sample size is $n = 50$ and the median equals 36.5. As with the median, the quartiles can easily be found from the stem-and-leaf plot of the data (Figure 3.4), which was

Stem	Leaf
0	8
1	1 1 5 7
2	2 4 5 6 6 6 6 7 7 8 9 9
3	0 1 3 3 4 5 5 6 7
4	0 0 3 5 6 6 6 7 7
5	1 1 1 5 6 8 9
6	1 5 6 6 9
7	0 3 9

The lower quartile is the median for the 25 observations below the median. It is the 13th smallest observation, or 27. The upper quartile is the median for the 25 observations above the median. It is the 13th largest observation, or 51.

In summary, since

$$\text{lower quartile} = 27, \ \text{median} = 36.5, \ \text{upper quartile} = 51,$$

roughly a quarter of the states had violent crime rates (i) below 27, (ii) between 27 and 36.5, (iii) between 36.5 and 51, and (iv) above 51. The distance between the upper quartile and the median, $51 - 36.5 = 14.5$, exceeds the distance $36.5 - 27 = 9.5$ between the lower quartile and the median. This commonly happens when the distribution is skewed to the right.

Software can easily find quartiles as well as other percentiles. In practice, percentiles other than the median are usually not reported for small data sets.

Measuring Variability: Interquartile Range

The difference between the upper and lower quartiles is called the ***interquartile range***, denoted by IQR. This measure describes the spread of the middle half of the observations. For the U.S. violent crime rates in Table 3.2, the interquartile range $\text{IQR} = 51 - 27 = 24$. The middle half of the murder rates fall within a range of 24. Like the range and standard deviation, the IQR increases as the variability increases,

and it is useful for comparing variability of different groups. For example, 12 years earlier in 1993, the quartiles of the U.S. statewide violent crime rates were 33 and 77, giving an IQR of $77 - 33 = 44$ and showing quite a bit more variability.

An advantage of the IQR over the ordinary range or the standard deviation is that it is not sensitive to outliers. The U.S. violent crime rates range from 8 to 79, so the range is 71. When we include the observation for D.C., which was 161, the IQR changes only from 24 to 28. By contrast, the range changes from 71 to 153.

For bell-shaped distributions, the distance from the mean to either quartile is about 2/3rd of a standard deviation. Then IQR is roughly $(4/3)s$. The insensitivity of the IQR to outliers has recently increased its popularity, although in practice the standard deviation is still much more common.

Box Plots: Graphing a Five-Number Summary of Positions

The median, the quartiles, and the maximum and minimum are five positions often used as a set to describe center and spread. For instance, software reports the following five-number summary for the violent crime rates (where Q1 = lower quartile, Q3 = upper quartile, regarding the median as the second quartile):

```
100% Max      79.0
 75% Q3       51.0
 50% Med      36.5
 25% Q1       27.0
  0% Min       8.0
```

The five-number summary provides a simple description of the data. It is the basis of a graphical display called the ***box plot*** that summarizes both the center and the variability. The *box* of a box plot contains the central 50% of the distribution, from the lower quartile to the upper quartile. The median is marked by a line drawn within the box. The lines extending from the box are called *whiskers*. These extend to the maximum and minimum, except for outliers, which are marked separately.

Figure 3.17 shows the box plot for the violent crime rates, in the format provided with SPSS software. The upper whisker and upper half of the central box are longer than the lower ones. This indicates that the right tail of the distribution, which corresponds to the relatively large values, is longer than the left tail. The plot reflects the skewness to the right of violent crime rates. (Some software also plots the mean on the box plot, representing it by a + sign.)

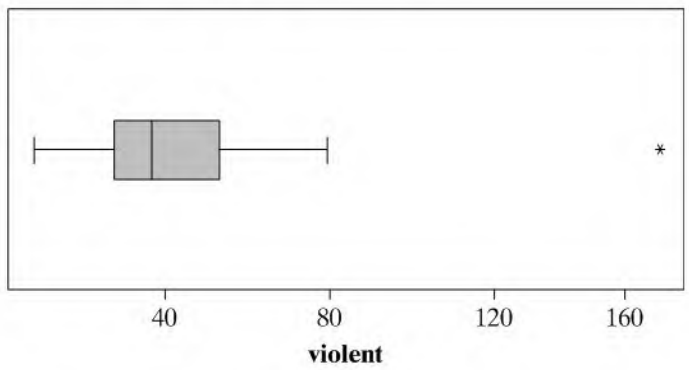


FIGURE 3.17: Box Plot of Violent Crime Rates of U.S. States and D.C.

Side-by-side box plots are useful for comparing two distributions. Figure 3.5 showed side-by-side stem-and-leaf plots of U.S. and Canadian murder rates. Figure 3.18 shows

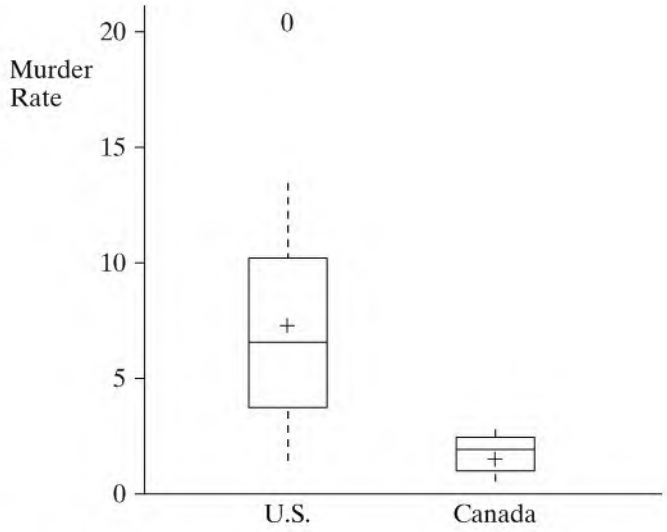


FIGURE 3.18: Box Plots for U.S. and Canadian Murder Rates

the side-by-side box plots. These side-by-side box plots reveal that the murder rates in the U.S. tend to be much higher and have much greater variability.

Outliers

Box plots identify outliers separately. To explain this, we now present a formal definition of an outlier.

Outlier

An observation is an ***outlier*** if it falls more than 1.5(IQR) above the upper quartile or more than 1.5(IQR) below the lower quartile.

In box plots, the whiskers extend to the smallest and largest observations only if those values are not outliers; that is, if they are no more than 1.5(IQR) beyond the quartiles. Otherwise, the whiskers extend to the most extreme observations within 1.5(IQR), and the outliers are marked separately. For instance, the statistical software SAS marks by an O (O for outlier) a value between 1.5 and 3.0(IQR) from the box and by an asterisk (*) a value even farther away.

Figure 3.18 shows one outlier for the U.S. with a very high murder rate. This is the murder rate of 20.3 (for Louisiana). For these data, the lower quartile = 3.9 and upper quartile = 10.3, so $IQR = 10.3 - 3.9 = 6.4$. Thus,

$$\text{Upper quartile} + 1.5(\text{IQR}) = 10.3 + 1.5(6.4) = 19.9.$$

Since $20.3 > 19.9$, the box plot highlights the observation of 20.3 as an outlier.

Why highlight outliers? It can be informative to investigate them. Was the observation perhaps incorrectly recorded? Was that subject fundamentally different from the others in some way? Often it makes sense to repeat a statistical analysis without an outlier, to make sure the conclusions are not overly sensitive to a single observation. Another reason to show outliers separately in a box plot is that they do not provide much information about the shape of the distribution, especially for large data sets.

In practice, the 1.5(IQR) criterion for an outlier is somewhat arbitrary. It is better to regard an observation satisfying this criterion as a *potential* outlier rather than a

definite outlier. When a distribution has a long right tail, some observations may fall more than 1.5 IQR above the upper quartile even if they are not separated far from the bulk of the data.

How Many Standard Deviations from the Mean? The *z*-Score

Another way to measure position is by the number of standard deviations that a point falls from the mean. For example, the U.S. murder rates shown in the box plot in Figure 3.18 have a mean of 7.3 and a standard deviation of 4.0. The value of 20.3 for Louisiana falls $20.3 - 7.3 = 13.0$ above the mean. Now, 13 is $13/4 = 3.25$ standard deviations. The Louisiana murder rate is 3.25 standard deviations above the mean.

The number of standard deviations that an observation falls from the mean is called its ***z*-score**. For the murder rates of Figure 3.18, Louisiana has a z-score of

$$z = \frac{20.3 - 7.3}{4.0} = \frac{\text{Observation} - \text{Mean}}{\text{Standard Deviation}} = 3.25.$$

By the Empirical Rule, for a bell-shaped distribution it is very unusual for an observation to fall more than three standard deviations from the mean. An alternative criterion regards an observation as an outlier if it has a z-score larger than 3 in absolute value. By this criterion, the murder rate for Louisiana is an outlier.

We'll study z-scores in more detail in the next chapter. We'll see they are especially useful for bell-shaped distributions.

3.5 BIVARIATE DESCRIPTIVE STATISTICS

In this chapter we've learned how to summarize categorical and quantitative variables graphically and numerically. In the next three chapters we'll learn about basic ideas of statistical inference for a categorical or quantitative variable. Most studies have more than one variable, however, and Chapters 7–16 present methods that can handle two or more variables at a time.

Association between Response and Explanatory Variables

With multivariable analyses, the main focus is on studying *associations* among the variables. There is said to be an ***association*** between two variables if certain values of one variable tend to go with certain values of the other.

For example, consider "religious affiliation," with categories (Protestant, Catholic, Other) and "ethnic group," with categories (Anglo-American, African-American, Hispanic). In the United States, Anglo-Americans are more likely to be Protestant than are Hispanics, who are overwhelmingly Catholic. African-Americans are even more likely to be Protestant. An association exists between religious affiliation and ethnic group, because the proportion of people having a particular religious affiliation changes as ethnic group changes.

An analysis of association between two variables is called a ***bivariate*** analysis, because there are two variables. Usually one is an outcome variable on which comparisons are made at levels of the other variable. The outcome variable is called the ***response variable***. The variable that defines the groups is called the ***explanatory variable***. The analysis studies how the outcome on the response variable *depends on* or is *explained by* the value of the explanatory variable. For example, when we describe how religious affiliation depends on ethnic group, religious affiliation is the response variable. In a comparison of men and women on income, income is the

response variable and gender is the explanatory variable. Income may depend on gender, not gender on income.

Often, the response variable is called the ***dependent variable*** and the explanatory variable is called the ***independent variable***. The terminology *dependent variable* refers to the goal of investigating the degree to which the response on that variable *depends on* the value of the other variable. We prefer not to use these terms, since *independent* and *dependent* are used for so many other things in statistical methods.

Comparing Two Groups Is a Bivariate Analysis

Chapter 7 will present descriptive and inferential methods for comparing two groups. For example, suppose we'd like to know whether men or women have more good friends, on the average. A GSS reports (for variable NUMFREND) that the mean number of good friends is 7.0 for men ($s = 8.4$) and 5.9 for women ($s = 6.0$). The two distributions have similar appearance, both being skewed to the right and with a median of 4.

Here, this is an analysis of two variables—number of good friends and gender. The response variable, number of good friends, is quantitative. The explanatory variable, gender, is categorical. In this case, it's common to compare means on the response variable for the categories of the categorical variable. Graphs are also useful, such as side-by-side box plots.

Bivariate Categorical Data

Chapter 8 will present methods for analyzing association between two categorical variables. Table 3.8 is an example of such data. This table results from answers to two questions on the 2006 General Social Survey. One asked whether homosexual relations are wrong. The other asked about the fundamentalism/liberalism of the respondent's religion. A table of this kind, called a ***contingency table***, displays the number of subjects observed at combinations of possible outcomes for the two variables. It displays how outcomes of a response variable are *contingent* on the category of the explanatory variable.

TABLE 3.8: Cross-Classification of Religion and Opinion about Homosexual Relations

	Opinion about Homosexual Relations				
Religion	Always Wrong	Almost Always Wrong	Sometimes Wrong	Not Wrong at All	Total
Fundamentalist	416	26	22	83	547
Liberal	213	29	52	292	586

Table 3.8 has eight possible combinations of responses. (Another possible outcome, "moderate" for the religion variable, is not shown here.) We could list the categories in a frequency distribution or construct a bar graph. Usually, though, it's more informative to do this for the categories of the response variable, separately for each category of the explanatory variable. For example, if we treat opinion about homosexual relations as the response variable, we could report the percentages in the four categories for homosexual relations, separately for each religious category.

Consider those who report being fundamentalist. Since 416/547 = 0.76, 76% believe homosexual relations are always wrong. Likewise, you can check that 5% believe they are almost always wrong, 4% believe they are sometimes wrong, and 15% believe they are not wrong at all. For those who report being liberal, since 213/586 = 0.36, 36% believe homosexual relations are always wrong. Likewise, you can check that 5% believe they are almost always wrong, 9% believe they are sometimes wrong, and 50% believe they are not wrong at all. There seems to be a definite association between opinion about homosexuality and religious beliefs, with religious fundamentalists being more negative about homosexuality. Chapter 8 will show many other ways of analyzing data of this sort.

Bivariate Quantitative Data

When both variables are quantitative, a plot we've not yet discussed is helpful. Figure 3.19 shows an example using the software SPSS to plot data from 38 nations on fertility (the mean number of children per adult woman) and the percentage of the adult population using cell phones. (The data are shown later in the text in Table 9.13.) Here, values of cell-phone use are plotted on the horizontal axis, called the ***x-axis***, and values of fertility are plotted on the vertical axis, called the ***y-axis***. The values of the two variables for any particular observation form a point relative to these axes. To portray graphically the sample data, we plot the 38 observations as 38 points. For example, the point at the top left of the plot represents Pakistan, which had a fertility of 6.2 children per woman but cell-phone use of only 3.5%. This graphical plot is called a ***scatterplot***.

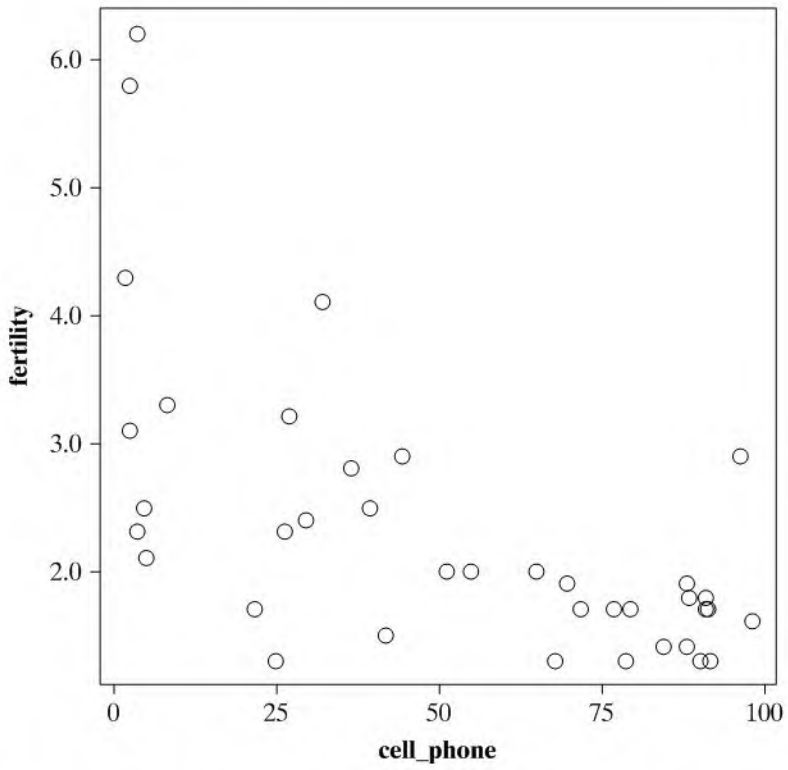


FIGURE 3.19: Scatterplot for Fertility and Percentage Using Cell Phones, for 38 Nations. The data are in Table 9.13 in Chapter 9.

The scatterplot shows a tendency for nations with higher cell-phone use to have lower levels of fertility. In Chapter 9 we'll learn about two ways to describe such as a trend. One way, called the ***correlation***, describes how strong the association is, in terms of how closely the data follow a *straight line trend*. For Figure 3.19, the correlation is -0.63. The negative value means that fertility tends to go *down* as cell-phone use goes *up*. By contrast, cell-phone use and GDP (gross domestic product, per capita) have a positive correlation of 0.83. As one goes up, the other also tends to go up.

The correlation takes values between -1 and $+1$. The larger it is in absolute value, that is, the farther from 0, the stronger the association. Cell-phone use is a bit more strongly associated with GDP than with fertility, because the correlation of 0.83 is larger in absolute value than the correlation of -0.63.

The second useful tool for describing the trend is ***regression analysis***. This provides a straight-line formula for predicting the value of the response variable from a given value of the explanatory variable. For Figure 3.19, this equation is

$$\text{Predicted fertility} = 3.4 - 0.02\,(\text{cell-phone use}).$$

For a country with no cell-phone use, the predicted fertility is $3.4 - 0.02(0) = 3.4$ children per mother. For a country with 100% of adults using cell phones, the predicted fertility is only $3.4 - 0.02(100) = 1.4$ children per mother.

Chapter 9 shows how to find the correlation and the regression line. Later chapters show how to extend the analysis to handle categorical as well as quantitative variables.

Analyzing More than Two Variables

This section has taken a quick look at analyzing associations between two variables. One important lesson from later in the text is that, *just because two variables have an association does not mean there is a causal connection*. For example, having more people in a nation using cell phones does not mean this is the reason the fertility rate is lower (for example, because people are talking on cell phones rather than doing what causes babies.) Perhaps high values on cell-phone use and low values on fertility are both a by-product of a nation being more economically advanced.

Most studies have *several* variables. The second half of this book (Chapters 10–16) shows how to conduct *multivariate* analyses. For example, to study what affects the number of good friends, we might want to simultaneously consider gender, age, whether married, educational level, whether attend religious services regularly, and whether live in urban or rural setting.

3.6 SAMPLE STATISTICS AND POPULATION PARAMETERS

Of the measures introduced in this chapter, the mean $\bar{y}$ is the most commonly reported measure of center and the standard deviation s is the most common measure of spread. We'll use them frequently in the rest of the text.

Since the values $\bar{y}$ and s depend on the sample selected, they vary in value from sample to sample. In this sense, they are variables. Their values are unknown before the sample is chosen. Once the sample is selected and they are computed, they become known sample statistics.

With inferential statistics, we shall distinguish between sample statistics and the corresponding measures for the population. Section 1.2 introduced the term *parameter* for a summary measure of the population. A statistic describes a sample, while a parameter describes the population from which the sample was taken. In this text, lowercase Greek letters usually denote population parameters and Roman letters denote the sample statistics.

Notation for Parameters

μ (Greek mu) and σ (Greek lowercase sigma) denote the mean and standard deviation of a variable for the population.

We call μ and σ the ***population mean*** and ***population standard deviation***. The population mean is the average of the observations for the entire population. The population standard deviation describes the variability of those observations about the population mean.

Whereas the statistics $\bar{y}$ and s are variables, with values depending on the sample chosen, the parameters μ and σ are constants. This is because μ and σ refer to just one particular group of observations, namely, the observations for the entire population. The parameter values are usually unknown, which is the reason for sampling and calculating sample statistics to estimate their values. Much of the rest of this text deals with ways of making inferences about unknown parameters (such as μ) using sample statistics (such as $\bar{y}$). Before studying these inferential methods, though, you need to learn some basic ideas of *probability*, which serves as the foundation for the methods. Probability is the subject of the next chapter.

3.7 CHAPTER SUMMARY

This chapter introduced ***descriptive statistics***—ways of *describing* data to summarize key characteristics of the data.

3.7.1 Overview of Tables and Graphs

- A ***frequency distribution*** summarizes the counts for possible values or intervals of values. A ***relative frequency*** distribution reports this information using percentages or proportions.
- A ***bar graph*** uses bars over possible values to portray a frequency distribution for a categorical variable. For a quantitative variable, a similar graphic is called a ***histogram***. It shows whether the distribution is approximately bell shaped, U shaped, skewed to the right (longer tail pointing to the right), or whatever.
- The ***stem-and-leaf plot*** is an alternative portrayal of data for a quantitative variable. It groups together observations having the same leading digit (stem), and shows also their final digit (leaf). For small samples, it displays the individual observations.
- The ***box plot*** portrays the quartiles, the extreme values, and any outliers. The box plot and the stem-and-leaf plot also can provide back-to-back comparisons of two groups.

Stem-and-leaf plots and box plots, simple as they are, are relatively recent innovations in statistics. They were introduced by the great statistician John Tukey (see Tukey 1977), who also introduced the terminology "software." See Cleveland (1994) and Tufte (2001) for other innovative ways to present data graphically.

3.7.2 Overview of Measures of Center

Measures of center describe the center of the data, in terms of a typical observation.

- The ***mean*** is the sum of the observations divided by the sample size. It is the center of gravity of the data.
- The ***median*** divides the ordered data set into two parts of equal numbers of observations, half below and half above that point.

- The lower quarter of the observations fall below the ***lower quartile***, and the upper quarter fall above the ***upper quartile***. These are the 25th and 75th ***percentiles***. The median is the 50th percentile. The quartiles and median split the data into four equal parts. They are less affected than the mean by outliers or extreme skew.
- The ***mode*** is the most commonly occurring value. It is valid for any type of data, though usually used with categorical data or discrete variables taking relatively few values.

3.7.3 Overview of Measures of Variability

Measures of variability describe the spread of the data.

- The ***range*** is the difference between the largest and smallest observations. The ***interquartile range*** is the range of the middle half of the data between the upper and lower quartiles. It is less affected by outliers.
- The ***variance*** averages the squared deviations about the mean. Its square root, the ***standard deviation***, is easier to interpret, describing a typical distance from the mean.
- The ***Empirical Rule*** states that for a bell-shaped distribution, about 68% of the observations fall within one standard deviation of the mean, about 95% fall within two standard deviations, and nearly all, if not all, fall within three standard deviations.

Table 3.9 summarizes the measures of center and variability. A ***statistic*** summarizes a sample. A ***parameter*** summarizes a population. ***Statistical inference*** uses statistics to make predictions about parameters.

TABLE 3.9: Summary of Measures of Center and Variability

Measure	Definition	Interpretation
Center		
Mean	$\bar{y} = \Sigma y_i/n$	Center of gravity
Median	Middle observation of ordered sample	50th percentile, splits sample into two equal parts
Mode	Most frequently occurring value	Most likely outcome, valid for all types of data
Variability		
Standard deviation	$s = \sqrt{\Sigma(y_i - \bar{y})^2/(n-1)}$	Empirical Rule: If bell shaped, 68%, 95% within $s, 2s$ of $\bar{y}$
Range	Difference between largest and smallest observation	Greater with more variability
Interquartile range	Difference between upper quartile (75th percentile) and lower quartile (25th percentile)	Encompasses middle half of data

3.7.4 Overview of Bivariate Descriptive Statistics

Bivariate statistics are used to analyze data on two variables together.

- Many studies analyze how the outcome on a ***response variable*** depends on the value of an explanatory variable.

- For categorical variables, a ***contingency table*** shows the number of observations at the combinations of possible outcomes for the two variables.
- For quantitative variables, a ***scatterplot*** graphs the observations, showing a point for each observation. The response variable is plotted on the y-axis and the explanatory variable is plotted on the x-axis.
- For quantitative variables, the ***correlation*** describes the strength of straight-line association. It falls between -1 and $+1$ and indicates whether the response variable tends to increase (positive correlation) or decrease (negative correlation) as the explanatory variable increases.
- A ***regression analysis*** provides a straight-line formula for predicting the value of the response variable using the explanatory variable. We study correlation and regression in detail in Chapter 9.

PROBLEMS

Practicing the Basics

3.1. Table 3.10 shows the number (in millions) of the foreign-born population of the United States in 2004, by place of birth.

(a) Construct a relative frequency distribution.
(b) Sketch the data in a bar graph.
(c) Is "Place of birth" quantitative or categorical?
(d) Use whichever of the following measures is relevant for these data: mean, median, mode.

TABLE 3.10

Place of Birth	Number
Europe	4.7
Asia	8.7
Caribbean	3.3
Central America	12.9
South America	2.1
Other	2.6
Total	34.3

Source: Statistical Abstract of the United States, 2006.

3.2. According to www.adherents.com, in 2006 the number of followers of the world's five largest religions were 2.1 billion for Christianity, 1.3 billion for Islam, 0.9 billion for Hinduism, 0.4 billion for Confucianism, and 0.4 billion for Buddhism.

(a) Construct a relative frequency distribution.
(b) Sketch a bar graph.
(c) Can you find a mean, median, or mode for these data? If so, do so and interpret.

3.3. A teacher shows her class the scores on the midterm exam in the stem-and-leaf plot:

```
6 | 5 8 8
7 | 0 1 1 3 6 7 7 9
8 | 1 2 2 3 3 3 4 6 7 7 7 8 9
9 | 0 1 1 2 3 4 4 5 8
```

(a) Identify the number of students and the minimum and maximum scores.
(b) Sketch a histogram with four intervals.

3.4. According to the *2005 American Community Survey*, in 2005 the United States had 30.1 million households with one person, 37.0 million with two persons, 17.8 million with three persons, 15.3 million with four persons, and 10.9 million with five or more persons.

(a) Construct a relative frequency distribution.
(b) Sketch a histogram. What is its shape?
(c) Report and interpret the (i) median, (ii) mode of household size.

3.5. Copy the "2005 statewide crime" data file from the text Web site (www.stat.ufl.edu/~aa/social/data.html). Use the variable, murder rate (per 100,000 population). In this exercise, do not use the observation for D.C. Using software,

(a) Construct a relative frequency distribution.
(b) Construct a histogram. How would you describe the shape of the distribution?
(c) Construct a stem-and-leaf plot. How does this plot compare to the histogram in (b)?

3.6. The OECD (Organization for Economic Cooperation and Development) consists of advanced, industrialized countries that accept the principles of representative democracy and a free market economy. Table 3.11 shows UN data for OECD nations on several variables: gross domestic product (GDP, per capita in U.S. dollars), percent unemployed, a measure of inequality based on comparing wealth of the richest 10% to the poorest 10%, public expenditure on health (as a percentage of the GDP), the number of physicians per 100,000 people, carbon dioxide emissions (per capita, in metric tons), the percentage of seats in parliament held by women, and female economic activity as

TABLE 3.11: UN Data for OECD Nations, Available as "OECD data" File at Text Web Site

Nation	GDP	Unemp.	Inequal.	Health	Physicians	C02	Women Parl.	Fem. Econ.
Australia	30,331	5.1	12.5	6.4	247	18	28.3	79
Austria	32,276	5.8	6.9	5.1	338	8.6	32.2	75
Belgium	31,096	8.4	8.2	6.3	449	8.3	35.7	72
Canada	31,263	6.8	9.4	6.9	214	17.9	24.3	83
Denmark	31,914	4.9	8.1	7.5	293	10.1	36.9	84
Finland	29,951	8.6	5.6	5.7	316	13	37.5	86
France	29,300	10.0	9.1	7.7	337	6.2	13.9	79
Germany	28,303	9.3	6.9	8.7	337	9.8	30.5	76
Greece	22,205	10.6	10.2	5.1	438	8.7	13	66
Iceland	33,051	2.5	..	8.8	362	7.6	33.3	87
Ireland	38,827	4.3	9.4	5.8	279	10.3	14.2	72
Italy	28,180	7.7	11.6	6.3	420	7.7	16.1	61
Japan	29,251	4.4	4.5	6.4	198	9.7	10.7	65
Luxembourg	69,961	4.6	..	6.2	266	22	23.3	68
Netherlands	31,789	6.2	9.2	6.1	315	8.7	34.2	76
New Zealand	23,413	3.6	12.5	6.3	237	8.8	32.2	81
Norway	38,454	4.6	6.1	8.6	313	9.9	37.9	87
Portugal	19,629	7.5	15	6.7	342	5.6	21.3	79
Spain	25,047	9.1	10.3	5.5	330	7.3	30.5	65
Sweden	29,541	5.6	6.2	8	328	5.9	45.3	87
Switzerland	33,040	4.1	9	6.7	361	5.6	24.8	79
United Kingdom	30,821	4.8	13.8	6.9	230	9.4	18.5	79
United States	39,676	5.1	15.9	6.8	256	19.8	15	81

Source: hdr.undp.org/statistics/data
Unemp. = % Unemployed, Inequal. = Measure of inequality, Women parl. = % of seats in parliament held by women, Fem. econ. = Female economic activity (% of male rate).

a percentage of the male rate. These data are the "OECD data" file at the text Web site.

(a) Construct a stem-and-leaf plot of the GDP values, by rounding and reporting the values in thousands of dollars (e.g., replacing $19,629 by 20).

(b) Construct a histogram corresponding to the stem-and-leaf plot in (a).

(c) Identify the outlier in each plot.

3.7. Recently, the statewide number of abortions per 1000 women 15 to 41 years of age, for states in the Pacific region of the United States, were: Washington, 26; Oregon, 17; California, 236; Alaska, 2; and Hawaii, 6 (*Statistical Abstract of the United States, 2006*).

(a) Find the mean.

(b) Find the median. Why is it so different from the mean?

3.8. Global warming seems largely a result of human activity that produces carbon dioxide emissions and other greenhouse gases. The *Human Development Report 2005*, published by the United Nations Development Programme, reported per capita emissions in 2002 for the eight largest countries in population size, in metric tons (1000 kilograms) per person: Bangladesh 0.3, Brazil 1.8, China 2.3, India 1.2, Indonesia 1.4, Pakistan 0.7, Russia 9.9, United States 20.1.

(a) For these eight values, find the mean and the median.

(b) Does any observation appear to be an outlier? Discuss its impact on how the mean compares to the median.

3.9. A Roper organization survey asked, "How far have environmental protection laws and regulations gone?" For the possible responses not far enough, about right, and too far, the percentages of responses were 51%, 33%, and 16%.

(a) Which response is the mode?

(b) Can you compute a mean or a median for these data? If so, do so; if not, explain why not.

3.10. A researcher in an alcoholism treatment center, to study the length of stay in the center for first-time patients, randomly selects ten records of individuals institutionalized within the previous two years. The lengths of stay, in days, were 11, 6, 20, 9, 13, 4, 39, 13, 44, and 7.

(a) Construct a stem-and-leaf plot.

(b) Find the mean and the standard deviation, and interpret.

(c) For a similar study 25 years ago, lengths of stay for ten sampled individuals were 32, 18,

55, 17, 24, 31, 20, 40, 24, 15. Compare results to those in the new study using (i) a back-to-back stem-and-leaf plot, (ii) the mean, (iii) the standard deviation. Interpret any differences you find.

(d) Actually, the new study also selected one other record. That patient is still institutionalized after 40 days. Thus, that patient's length of stay is at least 40 days, but the actual value is unknown. Can you calculate the mean or median for the complete sample of size 11 including this partial observation? Explain. (This observation is said to be *censored*, meaning that the observed value is "cut short" of its true, unknown value.)

3.11. Access the GSS at sda.berkeley.edu/GSS. Entering TVHOURS for the variable and year(2006) in the selection filter, you obtain data on hours per day of TV watching in the U.S. in 2006.

(a) Construct the relative frequency distribution for the values 0, 1, 2, 3, 4, 5, 6, 7 or more.

(b) How would you describe the shape of the distribution?

(c) Explain why the median is 2.

(d) The mean is larger than 2. Why do you think this is?

3.12. Table 3.12 shows 2003 female economic activity (number of women in labor force per 100 men in labor force), for countries in Western Europe. Construct a back-to-back stem-and-leaf plot of these values contrasted with those from South America in Table 3.4. What is your interpretation?

3.13. According to Statistics Canada, in 2000 household income in Canada had median $46,752 and mean $71,600. What would you predict about the shape of the distribution? Why?

3.14. Table 3.13 summarizes responses of 2333 subjects in the 2006 General Social Survey to the question, "About how often did you have sex during the last 12 months?"

(a) Report the median and the mode. Interpret.

(b) Treat this scale in a quantitative manner by assigning the scores 0, 0.1, 1.0, 2.5, 4.3, 10.8, and 17 to the categories, for approximate monthly frequency. Find the sample mean, and interpret.

TABLE 3.13

How Often Had Sex	Frequency
Not at all	595
Once or twice	205
About once a month	265
2 or 3 times a month	361
About once a week	343
2 or 3 times a week	430
More than 3 times a week	134

3.15. The 2004 GSS asked respondents "How often do you read the newspaper?" The possible responses were (every day, a few times a week, once a week, less than once a week, never), and the counts in those categories were (358, 222, 134, 121, 71).

(a) Identify the mode and the median response.

(b) Let y = number of times you read the newspaper in a week, measured as described above. For the scores (7, 3, 1, 0.5, 0) for the categories, find $\bar{y}$. How does it compare to the mean of 4.4 for the 1994 GSS?

3.16. According to the U.S. Bureau of the Census, *2005 American Community Survey*, the median earnings in the past 12 months was $32,168 for females and $41,965 for males, whereas the mean was $39,890 for females and $56,724 for males.

(a) Does this suggest that the distribution of income for each gender is symmetric, or skewed to the right, or skewed to the left? Explain.

(b) The results refer to 73.8 million females and 83.4 million males. Find the overall mean income.

TABLE 3.12

Country	Female Econ. Activity	Country	Female Econ. Activity	Country	Female Econ. Activity
Austria	66	Germany	71	Norway	86
Belgium	67	Greece	60	Portugal	72
Cyprus	63	Ireland	54	Spain	58
Denmark	85	Italy	60	Sweden	90
Finland	87	Luxembourg	58	U.K.	76
France	78	Netherlands	68		

Source: *Human Development Report, 2005*, United Nations Development Programme.

3.17. In 2003 in the United States, the median family income was \$55,800 for white families, \$34,400 for black families, and \$34,300 for Hispanic families (*Statistical Abstract of the United States, 2006*).
- **(a)** Identify the response variable and the explanatory variable for this analysis.
- **(b)** Is enough information given to find the median when all the data are combined from the three groups? Why or why not?
- **(c)** If the reported values were means, what else would you need to know to find the overall mean?

3.18. The GSS has asked, "During the past 12 months, how many people have you known personally that were victims of homicide." Table 3.14 shows a printout from analyzing responses.
- **(a)** Is the distribution bell shaped, skewed to the right, or skewed to the left?
- **(b)** Does the Empirical Rule apply to this distribution. Why or why not?
- **(c)** Report the median. If 500 observations shift from 0 to 6, how does the median change? What property does this illustrate for the median?

3.19. As of October 2006, an article in wikipedia.org on "Minimum wage" reported (in U.S. dollars) the minimum wage per hour for five nations: \$10.00 in Australia, \$10.25 in New Zealand, \$10.46 in France, \$10.01 in the U.K., \$5.15 in the U.S. Find the median, mean, range, and standard deviation **(a)** excluding the U.S., **(b)** for all five observations. Use the data to explain the effect of outliers on these measures.

3.20. *National Geographic Traveler* magazine recently presented data on the annual number of vacation days averaged by residents of eight different countries. They reported 42 days for Italy, 37 for France, 35 for Germany, 34 for Brazil, 28 for Britain, 26 for Canada, 25 for Japan, and 13 for the United States.
- **(a)** Find the mean and standard deviation. Interpret.
- **(b)** Report the five-number summary. (*Hint:* You can find the lower quartile by finding the median of the four values below the median.)

3.21. The Human Development Index (HDI) is an index the United Nations uses to give a summary rating for each nation based on life expectancy at birth, educational attainment, and income. In 2006, the ten nations (in order) with the highest HDI rating, followed in parentheses by the percentage of seats in their parliament held by women (which is a measure of gender empowerment) were Norway 38, Iceland 33, Australia 28, Ireland 14, Sweden 45, Canada 24, Japan 11, United States 15, Switzerland 25, Netherlands 34. Find the mean and standard deviation, and interpret.

3.22. The *Human Development Report 2006*, published by the United Nations (UN), showed life expectancies by country. For Western Europe, the values reported were

> Denmark 77, Portugal 77, Netherlands 78, Finland 78, Greece 78, Ireland 78, UK 78, Belgium 79, France 79, Germany 79, Norway 79, Italy 80, Spain 80, Sweden 80, Switzerland 80.

For Africa, the values reported (many of which were substantially lower than five years earlier because of the prevalence of AIDS) were

> Botswana 37, Zambia 37, Zimbabwe 37, Malawi 40, Angola 41, Nigeria 43, Rwanda 44, Uganda 47, Kenya 47, Mali 48, South Africa 49, Congo 52, Madagascar 55, Senegal 56, Sudan 56, Ghana 57.

- **(a)** Which group of life expectancies do you think has the larger standard deviation? Why?
- **(b)** Find the standard deviation for each group. Compare them to illustrate that s is larger for the group that shows more spread.

TABLE 3.14

VICTIMS	Frequency	Percent
0	1244	90.8
1	81	5.9
2	27	2.0
3	11	0.8
4	4	0.3
5	2	0.1
6	1	0.1

N	Mean	Std Dev	Max	Q3	Med	Q1	Min
1370	0.146	0.546	6	0	0	0	0

3.23. A report indicates that teacher's annual salaries in Ontario have a mean of \$50,000 and standard deviation of \$10,000 (Canadian dollars). Suppose the distribution has approximately a bell shape.
(a) Give an interval of values that contains about (i) 68%, (ii) 95%, (iii) all or nearly all salaries.
(b) Would a salary of \$100,000 be unusual? Why?

3.24. Excluding the U.S., the national mean number of holiday and vacation days in a year for OECD nations (see Exercise 3.6) is approximately bell shaped with a mean of 35 days and standard deviation of 3 days.[1]
(a) Use the Empirical Rule to describe the variability.
(b) The observation for the U.S. is 19. If this is included with the other observations, will the (i) mean increase, or decrease, (ii) standard deviation increase, or decrease?
(c) Using the mean and standard deviation for the other countries, how many standard deviations is the U.S. observation from the mean?

3.25. For GSS data on "the number of people you know who have committed suicide," 88.8% of the responses were 0, 8.8% were 1, and the other responses took higher values. The mean equals 0.145, and the standard deviation equals 0.457.
(a) What percentage of observations fall within one standard deviation of the mean?
(b) Is the Empirical Rule appropriate for the distribution of this variable? Why or why not?

3.26. The first exam in your Statistics course is graded on a scale of 0 to 100, and the mean is 76. Which value is most plausible for the standard deviation: -20, 0, 10, or 50? Why?

3.27. Grade point averages of graduating seniors at the University of Rochester must fall between 2.0 and 4.0. Consider the possible standard deviation values: -10.0, 0.0, 0.4, 1.5, 6.0.
(a) Which is the most realistic value? Why?
(b) Which value is *impossible*? Why?

3.28. According to the U.S. Census Bureau, the U.S. nationwide median selling price of homes sold in 2005 was \$184,100. Which of the following is the most plausible value for the standard deviation: **(a)** $-15{,}000$, **(b)** 1,000, **(c)** 10,000, **(d)** 60,000, **(e)** 1,000,000? Why?

3.29. For all homes in Gainesville, Florida, the residential electrical consumption[2] for the year 2006 had a mean of 10,449 and a standard deviation of 7489 kilowatt-hours (kWh). The maximum usage was 336,240 kWh.
(a) What shape do you expect this distribution to have? Why?
(b) Do you expect this distribution to have any outliers? Explain.

3.30. Residential water consumption (in thousands of gallons) in Gainesville, Florida in 2006 had a mean of 78 and a standard deviation of 119. What shape do you expect this distribution to have? Why?

3.31. According to *Statistical Abstract of the United States 2006*, mean salary (in dollars) of secondary school teachers in 2004 in the United States varied among states with a five-number summary of

```
100% Max      61,800 (Illinois)
 75% Q3       48,850
 50% Med      42,700
 25% Q1       39,250
  0% Min      33,100 (South Dakota)
```

(a) Find and interpret the range.
(b) Find and interpret the interquartile range.

3.32. Refer to the previous exercise.
(a) Sketch a box plot.
(b) Based on (a), predict the direction of skew for this distribution. Explain.
(c) If the distribution, although skewed, is approximately bell shaped, which value is most plausible for the standard deviation: (i) 100, (ii) 1000, (iii) 7000, (iv) 25,000? Explain.

3.33. Table 3.15 shows part of a computer printout for analyzing the murder rates (per 100,000) in the "2005 statewide crime" data file at the text Web site. The first column refers to the entire data set, and the second column deletes the observation for D.C. For each statistic reported, evaluate the effect of including the outlying observation for D.C.

3.34. During a recent semester at the University of Florida, computer usage[3] of students having accounts on a mainframe computer was summarized by a mean of 1921 and a standard deviation of 11,495 kilobytes of drive usage.
(a) Does the Empirical Rule apply to this distribution? Why?
(b) The five-number summary was minimum = 4, $Q1$ = 256, median = 530, $Q3$ = 1105, and maximum = 320,000. What does this suggest about the shape of the distribution? Why?
(c) Use the 1.5(IQR) criterion to determine if any outliers are present.

[1] *Source*: Table 8.9 in www.stateofworkingamerica.org, from The Economic Policy Institute.
[2] Data supplied by Todd Kamhoot, Gainesville Regional Utilities.
[3] Data supplied by Dr. Michael Conlon, University of Florida.

TABLE 3.15

Variable = MURDER			
N	51	N	50
Mean	5.6	Mean	4.8
Std Dev	6.05	Std Dev	2.57
	Quartiles		Quartiles
100% Max	44	100% Max	13
75% Q3	6	75% Q3	6
50% Med	5	50% Med	5
25% Q1	3	25% Q1	3
0% Min	1	0% Min	1
Range	43	Range	12
Q3-Q1	3	Q3-Q1	3
Mode	3	Mode	3

3.35. For each of the following, sketch what you expect a histogram to look like, and explain whether the mean or the median would be greater.
- **(a)** The selling price of new homes in 2008
- **(b)** The number of children ever born per woman age 40 or over
- **(c)** The score on an easy exam (mean = 88, standard deviation = 10, maximum possible = 100)
- **(d)** The number of cars owned per family
- **(e)** Number of months in which subject drove a car last year

3.36. For each of the following variables, indicate whether you would expect its relative frequency histogram to be bell shaped, U shaped, skewed to the right, or skewed to the left.
- **(a)** Exam score of easy exam, with mean = 88, standard deviation = 10, minimum = 65, lower quartile = 77, median = 85, upper quartile = 91, maximum = 100)
- **(b)** IQ for the general population
- **(c)** Number of times arrested in past year
- **(d)** Time needed to complete difficult exam (maximum time is 1 hour)
- **(e)** Age at death
- **(f)** Weekly church contribution (median is \$10 and mean is \$17)
- **(g)** Attitude toward legalization of abortion

3.37. For parts (a), (b), and (f) of the previous exercise, sketch box plots that would be plausible for the variable.

3.38. The January 2007 unemployment rates of the 27 countries in the European Union ranged from 3.2 (Denmark) to 12.6 (Poland), with lower quartile = 5.0, median = 6.7, upper quartile = 7.9, mean = 6.7, and standard deviation = 2.2. Sketch a box plot, labeling which of these values are used in the plot.

3.39. For the student survey data on number of times a week reading a newspaper, referred to in Exercise 1.11, Figure 3.20 shows a computer printout of the stem-and-leaf plot and the box plot.
- **(a)** From the box plot, identify the minimum, lower quartile, median, upper quartile, and maximum.
- **(b)** Identify these five numbers using the stem-and-leaf plot.
- **(c)** Do the data appear to contain any outliers? If so, identify.
- **(d)** The standard deviation is one of the following values—0.3, 3, 13, 23. Which do you think it is, and why?

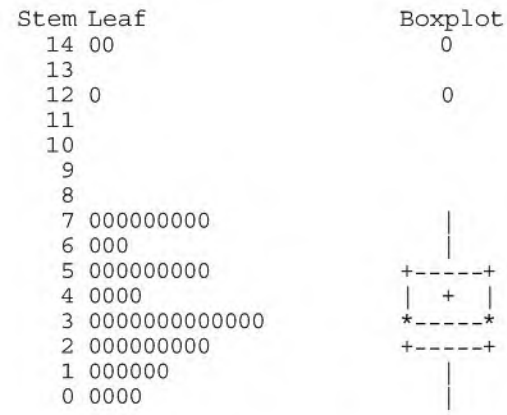


FIGURE 3.20

3.40. Infant mortality rates (number of infant deaths, per 1000 live births) are reported by the UN. In their 2006 report, the values for Africa had a five-number summary of

$$\text{min} = 54, \text{Q1} = 76, \text{median} = 81,$$
$$\text{Q3} = 101, \text{max} = 154.$$

The values for Western Europe had a five-number summary of

$$\text{min} = 3, \text{Q1} = 4, \text{median} = 4, \text{Q3} = 4, \text{max} = 5.$$

Sketch side-by-side box plots, and use them to describe differences between the distributions. (The plot for Europe shows that the quartiles, like the median, are less useful when the data are highly discrete.)

3.41. In 2004, the five-number summary for the statewide percentage of people without health insurance had a minimum of 8.9% (Minnesota), $Q1$ = 11.6, Med = 14.2, $Q3$ = 17.0, and maximum of 25.0% (Texas) (*Statistical Abstract of the United States, 2006*).
- **(a)** Sketch a box plot.

(b) Do you think that the distribution is symmetric, skewed to the right, or skewed to the left? Why?

3.42. High school graduation rates in the U.S. in 2004 had a minimum of 78.3 (Texas), lower quartile of 83.6, median of 87.2, upper quartile of 88.8, and maximum of 92.3 (Minnesota) (*Statistical Abstract of the United States, 2006*).

(a) Report and interpret the range and the interquartile range.

(b) Are there any outliers according to the 1.5(IQR) criterion?

3.43. Using software, analyze the murder rates from the "2005 statewide crime" data file at the text website.

(a) Using the data set without D.C., find the five-number summary.

(b) Construct a box plot, and interpret.

(c) Repeat the analyses, including the D.C. observation, and compare results.

3.44. A report by the OECD[4] indicated that annual water consumption for nations in the OECD (see Exercise 3.6) was skewed to the right, with values (in cubic meters per capita) having a median of about 500 and ranging from about 200 in Denmark to 1700 in the U.S. Consider the possible values for the IQR: -10, 0, 10, 350, 1500. Which is the most realistic value? Why?

3.45. According to values from the *Human Development Report*, published by the United Nations (hdr.undp.org), carbon dioxide emissions in 2005 for the 25 nations in the European Union (EU) as of 2005 had a mean of 8.3 and standard deviation of 3.3, in metric tons per capita. All values were below 12, except Luxembourg which had a value of 21.1.

(a) How many standard deviations above the mean was the value for Luxembourg?

(b) Sweden's observation was 5.8. How many standard deviations below the mean was it?

(c) The carbon dioxide emissions were 16.5 for Canada and 20.1 for the U.S. Relative to the distribution for the EU, find and interpret the z-score for (i) Canada, (ii) the U.S.

3.46. The United Nations publication *Energy Statistics Yearbook* (unstats.un.org/unsd/energy) lists consumption of energy. For the 25 nations that made up the EU in 2006, the energy values (in kilograms per capita) had a mean of 4998 and a standard deviation of 1786.

(a) Italy had a value of 4222. How many standard deviations from the mean was it?

(b) The value for the U.S. was 11,067. Relative to the distribution for the EU, find its z-score. Interpret.

(c) If the distribution of EU energy values were bell shaped, would a value of 11,067 be unusually high? Why?

3.47. A study compares Democrats and Republicans on their opinions about national health insurance (favor or oppose).

(a) Identify the response variable and the explanatory variable.

(b) Explain how the data could be summarized in a contingency table.

3.48. Table 3.16 shows reported happiness for those subjects in the 2004 GSS who said that they attend religious services rarely or frequently (variables ATTEND and HAPPY).

(a) Identify the response variable and the explanatory variable.

(b) At each level of religious attendance, find the percentage who reported being very happy.

(c) Does there seem to be an association between these variables? Why?

3.49. For recent United Nations data for several nations, a prediction equation relating fertility (the mean number of children per adult woman) and percentage of people using the Internet is

$$\text{Predicted fertility} = 3.2 - 0.04\,(\text{Internet use}).$$

(a) Compare the predicted fertility of a nation with 50% use of the Internet (the United States) to a nation with 0% use (Yemen).

(b) The correlation is -0.55. Explain what the negative value represents.

3.50. Refer to the previous exercise. A prediction equation relating fertility and percentage of people using contraceptive methods is:

$$\text{Predicted fertility} = 6.6 - 0.065\,(\text{contraceptive use})$$

and the correlation is -0.89.

TABLE 3.16

	Happiness			
Religious Attendance	Very Happy	Pretty Happy	Not Too Happy	Total
Nearly every week or more	200	220	29	449
Never or less than once a year	72	185	53	310

[4] *OECD Key Environmental Indicators 2005.*

(a) What type of pattern would you expect for the points in a scatterplot for these data?

(b) Which variable seems to be more strongly associated with fertility—Internet use or contraceptive use? Why?

3.51. For the data for OECD nations in Table 3.11 in Exercise 3.6, use software to construct a scatterplot relating y = carbon dioxide emissions and x = GDP.

(a) Based on this plot, would you expect the correlation between these variables to be positive or negative? Why?

(b) Do you see an observation that falls apart from the others? Identify the nation.

3.52. Refer to the previous exercise. The correlation with carbon dioxide emissions is 0.03 for female economic activity and -0.52 with number of physicians. Which variable is more strongly associated with carbon dioxide emissions? Why?

3.53. What is the difference between the descriptive measures symbolized by

(a) $\bar{y}$ and μ?

(b) s and σ?

Concepts and Applications

3.54. For the "Student survey" data file at the text Web site (see Exercise 1.11 on page 8), use software to conduct graphical and numerical summaries for

(a) distance from home town,

(b) weekly hours of TV watching. Describe the shapes of the distributions, and summarize your findings.

3.55. Refer to the data file your class created for Exercise 1.12 (page 9). For variables chosen by your instructor, conduct descriptive statistical analyses. In your report, give an example of a research question that could be addressed using your analyses, identifying response and explanatory variables. Summarize and interpret your findings.

3.56. Table 3.17 shows annual gun death rates (including homicide, suicide, and accidental deaths) per 100,000 population in advanced industrialized nations. Prepare a report in which you summarize the data using graphical and numerical methods from this chapter.

3.57. For the "2005 statewide crime" dataset at the text Web site, consider violent crime rate and percentage with income below the poverty level. Pose a research question for these variables relating to the direction of their association, identifying the response variable and explanatory variable. Using software, construct a scatterplot and find the correlation. Interpret, and indicate what the scatterplot and correlation suggest about the research question.

3.58. Refer to Exercise 3.6. Pose a research question relating to the correlation between public expenditure on health and the number of physicians per 100,000 people. Using software, analyze data in Table 3.11 to address this question, and summarize your analyses and conclusions.

3.59. Zagat restaurant guides publish ratings of restaurants for many large cities around the world (see www.zagat.com). The review for each restaurant gives a verbal summary as well as a 0-to-30-point rating of the quality of food, decor, service and the cost of a dinner with one drink and tip. Figure 3.21 shows side-by-side box plots of the cost for Italian restaurants in Boston, London, and New York (Little Italy and Greenwich Village neighborhoods). Summarize what you learn from these plots.

3.60. Refer to the previous exercise. The data are available in the "Zagat data" file at the text Web site. For the 83 restaurants listed in London, the quality of food rating has a correlation of 0.61 with decor rating, 0.81 with service rating, and 0.53 with cost rating. Summarize what you learn from these correlations.

3.61. Exercise 3.21 introduced the Human Development Index (HDI). Go to hdr.undp.org/statistics/data/ and get the latest HDI ratings for Sub-Saharan African nations and separately for Westernized

TABLE 3.17

Nation	Gun Deaths	Nation	Gun Deaths	Nation	Gun Deaths
Australia	1.7	Greece	1.8	Norway	2.6
Austria	3.6	Iceland	2.7	Portugal	2.1
Belgium	3.7	Ireland	1.5	Spain	0.7
Canada	3.1	Italy	2.0	Sweden	2.1
Denmark	1.8	Japan	0.1	Switzerland	6.2
Finland	4.4	Luxembourg	1.9	U.K.	0.3
France	4.9	Netherlands	0.8	U.S.	9.4
Germany	1.5	New Zealand	2.3		

Source: Small Arms Survey, Geneva, 2007.

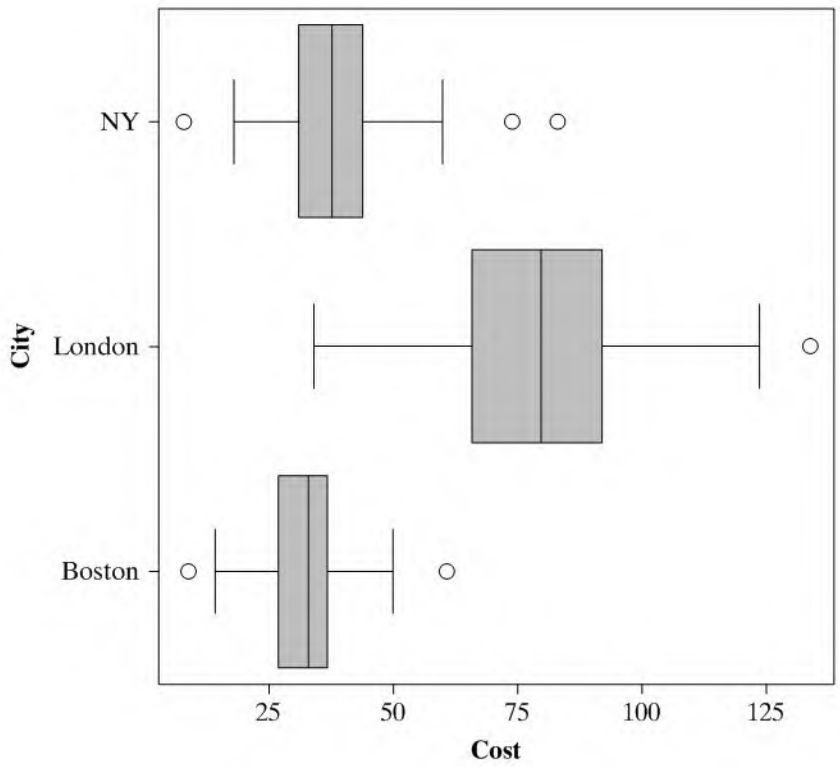


FIGURE 3.21

nations listed at the site as "High-income OECD." (One way to do this is to click on "Access a wide range of data tools" and then "Build table" and make the appropriate choices.) Using graphical and numerical methods of this chapter, summarize the data.

3.62. The incomes of players on the New York Yankees baseball team in 2006 can be summarized by the numbers[5] $2,925,000 and $7,095,078. One of these was the median and one was the mean. Which value do you think was the mean? Why?

3.63. In 2001, the U.S. Federal Reserve sampled about 4000 households to estimate overall net worth of a family. The Reserve reported the summaries $86,100 and $395,500. One of these was the mean, and one was the median. Which do you think was the median? Why?

3.64. A U.S. Federal Reserve study in 2000 indicated that for those families with annual incomes above $100,000, their median net worth was about $500,000 both in 1995 and in 1998, but their mean net worth rose from $1.4 million in 1995 to $1.7 million in 1998. A newspaper story about this said that the mean uses "a calculation that captures the huge gains made by the wealthiest Americans." Why would the median not necessarily capture those gains?

3.65. The fertility rate (mean number of children per adult woman) varies in Western European countries between a low of 1.3 (Italy and Spain) and a high of 1.9 (Ireland). For each woman, the number of children is a whole number, such as 0 or 1 or 2. Explain why it makes sense to measure a mean number of children per adult woman (which is not a whole number); for example, to compare these rates among European countries or with Canada (1.5), the U.S. (2.0), and Mexico (2.4).

3.66. According to a report from the U.S. National Center for Health Statistics, for males age 25–34 years, 2% of their heights are 64 inches or less, 8% are 66 inches or less, 27% are 68 inches or less, 39% are 69 inches or less, 54% are 70 inches or less, 68% are 71 inches or less, 80% are 72 inches or less, 93% are 74 inches or less, and 98% are 76 inches or less. These are called *cumulative percentages*.

(a) Find the median male height.

(b) Nearly all the heights fall between 60 and 80 inches, with fewer than 1% falling outside that range. If the heights are approximately bell shaped, give a rough approximation for the standard deviation. Explain your reasoning.

3.67. Give an example of a variable for which the mode applies, but not the mean or median.

3.68. Give an example of a variable having a distribution that you expect to be

(a) approximately symmetric,

(b) skewed to the right,

[5]http://usatoday.com/sports/baseball/salaries/

(c) skewed to the left,
(d) bimodal,
(e) skewed to the right, with a mode and median of 0 but a positive mean.

3.69. To measure center, why is the
(a) median sometimes preferred over the mean?
(b) Mean sometimes preferred over the median? In each case, give an example to illustrate your answer.

3.70. To measure variability, why is
(a) The standard deviation s usually preferred over the range?
(b) The IQR sometimes preferred to s?

3.71. Answer true or false to the following:
(a) The mean, median, and mode can never all be the same.
(b) The mean is always one of the data points.
(c) The median is the same as the second quartile and the 50th percentile.
(d) For 67 sentences for murder recently imposed using U.S. Sentencing Commission guidelines, the median length was 160 months and the mean was 251 months. This distribution is probably skewed to the right.

For multiple-choice problems 3.72–3.74, select the best response.

3.72. In Canada, based on the 2001 census, for the categories (Catholic, Protestant, Other Christian, Muslim, Jewish, None, Other) for religious affiliation, the relative frequencies were (42%, 28%, 4%, 2%, 1%, 16%, 7%) (*Statistics Canada*).
(a) The median religion is Protestant.
(b) Only 2.7% of the subjects fall within one standard deviation of the mean.
(c) The mode is Catholic.
(d) The Jewish response is an outlier.

3.73. The 2004 GSS asked whether having sex before marriage is (always wrong, almost always wrong, wrong only sometimes, not wrong at all). The response counts in these four categories were (238, 79, 157, 409). This distribution is
(a) Skewed to the right
(b) Approximately bell shaped
(c) Bimodal
(d) Shape does not make sense, since the variable is nominal

3.74. In a study of graduate students who took the Graduate Record Exam (GRE), the Educational Testing Service reported that for the quantitative exam, U.S. citizens had a mean of 529 and standard deviation of 127, whereas the non-U.S. citizens had a mean of 649 and standard deviation of 129.
(a) Both groups had about the same amount of variability in their scores, but non-U.S. citizens performed better, on the average, than U.S. citizens.
(b) If the distribution of scores was approximately bell shaped, then almost no U.S. citizens scored below 400.
(c) If the scores range between 200 and 800, then probably the scores for non-U.S. citizens were symmetric and bell shaped.
(d) A non-U.S. citizen who scored three standard deviations below the mean had a score of 200.

3.75. A teacher summarizes grades on the midterm exam by

$$\begin{aligned}&\text{Min} = 26, \text{Q1} = 67, \text{Median} = 80,\\ &\text{Q3} = 87, \text{Max} = 100,\\ &\text{Mean} = 76, \text{Mode} = 100,\\ &\text{Standard dev.} = 76, \text{IQR} = 20.\end{aligned}$$

She incorrectly recorded one of these. Which one do you think it was? Why?

3.76. Ten people are randomly selected in Florida and another ten people are randomly selected in Alabama. Table 3.18 provides summary information on mean income. The mean is higher in Alabama both in rural areas and in urban areas. Which state has the larger overall mean income? (The reason for this apparent paradox is that mean urban incomes are larger than mean rural incomes for both states and the Florida sample has a higher proportion of urban residents.)

TABLE 3.18

State	Rural		Urban	
Florida	\$26,000	($n = 3$)	\$39,000	($n = 7$)
Alabama	\$27,000	($n = 8$)	\$40,000	($n = 2$)

3.77. Refer to Table 3.2 (page 34). Explain why the mean of these 50 observations is not necessarily the same as the violent crime rate for the entire U.S. population.

3.78. For a sample with mean $\bar{y}$, adding a constant c to each observation changes the mean to $\bar{y} + c$, and the standard deviation s is unchanged. Multiplying each observation by c changes the mean to $c\bar{y}$ and the standard deviation to $|c|s$.
(a) Scores on a difficult exam have a mean of 57 and a standard deviation of 20. The teacher boosts all the scores by 20 points before awarding grades. Report the mean and standard deviation of the boosted scores.
(b) Suppose that annual income of Canadian lawyers has a mean of \$100,000 and a standard deviation of \$30,000. Values are converted to

British pounds for presentation to a British audience. If one British pound equals $2.00, report the mean and standard deviation in British currency.

(c) Observations from a survey that asks about the number of miles travelled each day on mass transit are to be converted to kilometer units (1 mile = 1.6 kilometers). Explain how to find the mean and standard deviation of the converted observations.

***3.79.** Show that $\Sigma(y_i - \bar{y})$ must equal 0 for any collection of observations $y_1, y_2, \ldots, y_n$.

***3.80.** The Russian mathematician Tchebysheff proved that for any $k > 1$, the proportion of observations that fall more than k standard deviations from the mean can be no greater than $1/k^2$. This holds for *any* distribution, not just bell-shaped ones.

(a) Find the upper bound for the proportion of observations falling (i) more than two standard deviations from the mean, (ii) more than three standard deviations from the mean, (iii) more than ten standard deviations from the mean.

(b) Compare the upper bound for $k = 2$ to the approximate proportion falling more than two standard deviations from the mean in a bell-shaped distribution. Why is there a difference?

3.81.** The ***least squares property of the mean states that the data fall closer to $\bar{y}$ than to any other number c, in the sense that the sum of squares of deviations of the data about their mean is smaller than the sum of squares of their deviations about c. That is,

$$\sum(y_i - \bar{y})^2 < \sum(y_i - c)^2.$$

If you have studied calculus, prove this property by treating $f(c) = \Sigma(y_i - c)^2$ as a function of c and deriving the value of c that provides a minimum. (*Hint*: Take the derivative of $f(c)$ with respect to c and set it equal to zero.)

This page intentionally left blank