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.

1.4 Cl

Practi
1.1.

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%

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.

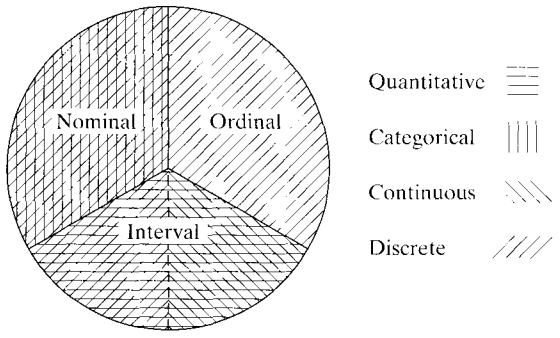
For quantitative variables the distinction between discrete and continuous variables can be blurry, because of how variables are actually measured. In practice, we round continuous variables when measuring them, so the measurement is actually discrete. We say that an individual is 21 years old whenever 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