PART I

Univariate Description

COPYRIGHTED MATERIAL

Chapter 1

USING STATISTICS

LEARNING OBJECTIVES

When you have finished this chapter, you will be able to:

- Distinguish the uses of statistics for description, inference, hypothesis testing, and prediction.
- Distinguish between the exploratory and confirmatory uses of statistics.
- Understand the distinction between sample statistics and population parameters.
- Understand how statistics are related to the process of constructing and verifying theory.
- Know how to write a report using statistical information.
- Identify units of analysis and cases in research problems.
- Understand variables and apply the notion of levels of measurement.

WHY STUDY STATISTICS?

Whether you wish it or not, you are a consumer of statistics. The news media present a constant flow of information, with government officials and pundits often throwing out number after number, ranging from current unemployment figures to opinion poll results about how some administration is doing in office. During football season, colleges are ranked based on polling coaches and feeding statistical data into computers. In newspapers, leading economic indicators, corporate performance and profits, results of studies both social and medical, car sales, and weather forecasts are all presented to the public in statistical language.

How does one process information and cope with the relentless barrage of statistical data without becoming overwhelmed? And, perhaps even more importantly, how does one become an intelligent and critical user of statistical information? Statistics are used not only to inform us, but to influence us. As a common aphorism puts it: "Statistics don't lie, but liars use statistics."

For a student of statistics, the answer lies in becoming an informed and discerning user of numerical information. The job of a student of statistics is to derive useful and usable information from a flood of data too large and complicated to be understood without being summarized. Learning to discern what is essential and what is not in quantitative data is an art and craft that requires practice and experience. But it is cultivated through careful learning. Some people devote their entire careers to accomplishing this and becoming experts—for example, university professors. Unless he or she is a statistics major, an undergraduate student does not have the time to devote a career to learning statistics in the four years of study for the BA or BS degree. Nevertheless, the crucial tools of becoming an informed user of statistical data can be learned in an introductory-level course in statistics in the time span of a quarter or a semester.

This book has been written to guide you through your first course in statistics at the undergraduate level. The only assumption that the authors have made is that everyone is capable of learning statistics. Accordingly, the book is not a textbook in mathematics. Formulas have been kept to a minimum. Where their use is unavoidable, the authors have gone to great lengths to explain every component and guide you through their application. Mathematical proofs are deliberately avoided throughout the book. Students who wish to learn proofs and acquire higher levels of statistics knowledge will find many courses on their campus that will satisfy their intellectual curiosity. The goal of the book is to help you learn and become an informed user of statistics whether your eventual plan is to obtain a degree and get employed or to undertake graduate work.

TASKS FOR STATISTICS: DESCRIBING, INFERRING, TESTING, PREDICTING

A first step in learning statistics is to appreciate the human capacity to summarize information through visual inspection and numerical indicators. Graphs and other charts can be used to summarize lots of information that would require many pages if each piece of information were written down without graphs. Through visual inspection, the statistician can begin to look for patterns or trends in data to explore the possibility of uncovering relationships. In addition to visual inspection, the statistician uses simple arithmetic to compute single numbers (indexes or numerical indicators) that summarize data and buttress results derived from visual gleaning.

Having described the data in summary form, the next major goal of the statistician is to draw inferences or arrive at conclusions using simple mathematical calculations. In an increasingly complex world, rarely do researchers use the entire population or universe of cases to conduct a study. Instead, they use a subset of the population. This subset is called a sample. For instance, a researcher wishing to study risk factors for divorce in the United States will most likely not have time, money, and resources to study the entire population of couples who married at some point and divorced at a later time. In research, what frequently happens is that a researcher obtains a representative sample of couples that he or she studies.

While every research project is unique, we can summarize the four main ways that statistics are used in social research.

Describing Variation and Covariation

Descriptive statistics are tools that are used to summarize and index various things about the variation in the distribution of the scores of cases on one (or more) variable(s). This is one of our main tasks: to summarize large amounts of information in such a way that we can quickly, accurately, and honestly communicate about the main patterns that are present in the data.

There are a huge number of descriptive statistical indexes and visualizations. In this book, we provide only an introduction. Underlying all statistical description are a few—and only a few—very important ideas. If you grasp these ideas, you may always consult a textbook on the technical details of their application. In the first portion of the book we focus on these core ideas: the frequency distribution, central tendency, dispersion, and distributional shape. Later on, we will introduce the other core concept of descriptive statistics: association or covariation.

Inferring to the Population

In most social science applications of statistics, we observe only a ***sample*** of cases from a larger ***population***. We talk *to* some students, but would like to talk *about* all students. We talk *to* some citizens in a survey, but we want to make a generalization *about* all citizens. The second major task of statistics is to make inferences about variation and covariation in a population, based on the information available from a sample drawn from that population.

A number that summarizes variation or covariation in a sample is referred to as a ***statistic.*** If a number is calculated using the population, it is described as a ***parameter.*** For example, suppose researchers take a sample from the 2010 United States Census of Population, calculate the percentage of married individuals, and arrive at a figure of 50 percent; this would be a statistic, since they did not use the entire census. However, if they do the same computations on the full census and report a figure of 47 percent, in this case the 47 percent constitutes a parameter, since it was based on the entire population.

There are two aspects to this task: making an estimate of the value of some statistic and assessing how confident we are about the estimate. Estimating the value of a population parameter (e.g., the proportion of all Americans who are currently married) based on sample information (e.g., the proportion of the respondents to the General Social Survey [GSS] of the National Opinion Research Center [NORC] at the University of Chicago who say that they are married) is often quite straightforward for basic statistics. Determining how confident we are in making that inference is a bit more complicated, and is the main subject of ***inferential statistics.***

Almost all the statistical procedures we discuss in this text assume that, when we have selected a sample of cases from a defined population, we have used *probability sampling methods.* That is, we can state precisely what the likelihood is that a given case was selected from the population to be in the sample. If we select one student at random from a class of 10, the probability of any one student being selected is exactly 0.10.

CHECK QUIZ 1.1

1. Statistics are to parameters as
   a. samples are to populations.
   b. populations are to samples.
   c. medians are to standard deviations.
   d. standard deviations are to medians.
2. The goal of all techniques for selecting probability samples is to select samples that are
   a. very large.
   b. nonrandom.
   c. easily located.
   d. representative.
3. We use inferential statistics
   a. to make generalizations to a sample from a population.
   b. when we are working with data from a census.
   c. when we are using nonprobability sampling methods.
   d. to make generalizations to a population from a sample.

Probability sampling methods are often used in social science research to try to ensure that a sample is representative of the population from which it was drawn. Not all social science research, however, uses probability sampling (and there are often very good reasons for not doing so). While we may use descriptive statistics to describe variation and covariation in nonprobability samples, we mostly cannot use inferential statistics with them.

Testing Hypotheses

The third major task of statistics is to test hypotheses. As we will see shortly, statistical analysis in the social sciences is usually part of a larger intellectual project to develop and verify theories—explanations of how the social world works that go beyond a particular sample or even a particular population.

Descriptive statistics are often used to explore a set of observations and make effective summaries or empirical generalizations. That is, one way to do statistics is to start with the data and move toward more abstract and general statements. ***Hypothesis testing***, however, uses a different logic. It begins with a speculation about how the social world works (based on prior research, observation, and previous empirical work). Using this prior theory, we logically deduce what we should observe in a particular sample if the theory is correct, and what we might observe if it is not correct. That is, we make a hypothesis about the values of sample statistics in advance, and then we evaluate whether the evidence from the sample is consistent with our expectations.

Hypothesis testing uses both descriptive and inferential statistics to assess whether a set of observations is consistent with a theory. Rather than an *exploratory* application of statistics, it is a *confirmatory* application; that is, we seek to (dis)confirm theories using sample observations. Note that a characteristic of a sample is typically called a statistic, and a characteristic of a population from which that sample was derived is called a parameter. In hypothesis testing, we use sample statistics (characteristics) to arrive at best guesses or estimates of population parameters.

Predicting

In addition to describing, inferring, and hypothesis testing, statistical techniques are used for making predictions. In the use of statistics for ***prediction***, we create a *model* that describes our knowledge of how one or more causes (or predictors or independent variables) produce an outcome (or dependent variable). We then use the model to describe the patterns that we see in our sample. We evaluate whether our model does a good job or a poor job of describing the sample data. If the *goodness of fit* is adequate, we can then use the model to make predictions—either about the individual cases in the sample or with regard to new cases.

Models for statistical prediction, like the regression models that we examine near the end of this text, apply description, inference, and hypothesis-testing tools. The goal of the analysis, however, is to create a formal and mathematical description of social processes (the model), and see what this model can (and can't) help us explain about our current observations and predict about new observations that we might make.

Each of the four main tasks of statistics builds on the previous one. First we describe and explore observations in a sample. Then we infer how confident we are that the sample may generalize to a population. Then we may use statistics to test whether theories are consistent with our observations. Finally, we may use statistics to build predictive models of the social processes that produce variation and covariation.

CHECK QUIZ 1.2

1. The statement that "49 percent of the state's likely voters support Melvin, with a 3 percent margin of error" corresponds to which of the tasks of statistics?
   a. describing variation and covariation
   b. inference to the population
   c. testing a hypothesis
   d. prediction
2. The statement that "grades in statistics courses are directly proportional to the amount of time spent studying statistics" corresponds to which of the four tasks of statistics?
   a. describing variation and covariation
   b. inference to the population
   c. testing a hypothesis
   d. prediction
3. The statement that "contrary to our prediction, we found no association between the volume of chocolate consumed and one's grade on the midterm" corresponds to which of the four tasks of statistics?
   a. describing variation and covariation
   b. inference to the population
   c. testing a hypothesis
   d. prediction
4. The statement that "scores on the midterm ranged from a low of 12 to a high of 14" corresponds to which of the four tasks of statistics?
   a. describing variation and covariation
   b. inference to the population
   c. testing a hypothesis
   d. prediction

STATISTICS IN THE RESEARCH PROCESS

Statistics are used for describing, inferring, hypothesis testing, and prediction in almost every area of modern society. Businesses, government agencies, and nongovernmental organizations all collect and analyze data to understand patterns and trends. Scientific researchers do many of the same things using statistical techniques. But, in addition, research scientists use statistics as part of the process of developing and testing theories.

If you are reading this book, you are probably taking a basic course in statistics as part of your introduction to one of the social sciences. The fundamental goal of the social sciences is to create theories that help us to understand and explain the social world. Statistics play a key role in this enterprise. In this section, we briefly examine the ways that statistics fit in the research process. Then, we show what this looks like in practice—how research scientists apply statistics to problems and report their results.

Theories and Social Research

In social research, the use of statistics is never an end in itself, but rather a means to achieving an end. That end is the construction and validation of theory.

A theory is simply a proposed explanation or story about why something is the way it is, or works the way it does. Everyone develops theories all the time, in this very general sense of the word. So what makes a social-scientific theory different from any commonsense explanation? One important distinction is that a scientific theory is a *proposed* explanation. Scientific theories are never *proven* to be true. Theories can be useful or not; they can be shown by evidence to be consistent with particular observations or not. But theories are always tentative explanations that might apply to help us understand something. Another important distinction between a scientific theory and some other kinds of explanations is that scientific theories can be evaluated as useful or not by observation following intersubjective methods. That is, any two people following the same rules of observation and analysis can arrive at the same conclusion about whether a theory is a useful explanation of some phenomenon. This is part of why you are studying statistics. Statistics are part of the tool kit of shared methodologies that scientists use to connect theory and observation.

Scientific theories are general or abstract statements that provide proposed explanations for whole classes of specific phenomena. Our observations of the social world are much more specific and narrow, so there is always a gap between an abstract general explanation of a whole population or class of phenomena and our particular observations. While there are many issues in making valid connections between abstract theories and concrete observations, statistical tools address some of the issues of inference from samples to more general populations.

Social science research can begin with abstract theory, but most often it begins with an observation of a particular phenomenon in a particular place and time. That is, most social science research begins with *observing a datum.* To try to make sense of our observation, we always make comparisons—we may observe the same phenomenon over time, we may compare it to other phenomena that are similar to it, or we may compare the phenomenon to others that differ from it. We *form empirical generalizations* by making comparisons. That is, we see patterns, similarities, and differences.

The next step, having made some empirical generalization, is to develop a tentative explanation of it—a theory. We may draw on previous theories of the same phenomenon or theories about other phenomena that seem similar, or we may imagine wholly new concepts and ways of thinking about our observations. There are only three rules that the scientist needs to follow at this step: the theory we develop must apply to a general class of phenomena, we must be able to devise ways in which we can share an intersubjective understanding of the phenomenon, and it must be possible for us to prove that our theory fails or is not useful in explaining the phenomenon.

Once we have a proposed theory, we develop a *research hypothesis.* A hypothesis in research is a statement of an expected relationship between variables. If a researcher has a hunch that people with more education tend to have more income, the hypothesis implied here is that the higher the educational attainment, the higher the income. The researcher must first place this hypothesis within a broader scheme as to why and how those with higher educational attainment might earn more income. In other words, the researcher has to come up with some plausible explanations that elucidate the expected relationship between education and income. The researcher might argue, for example, that education is a form of investment that people make in themselves. Formal education through schooling takes a long time, but during that process, individuals learn skills and knowledge that can later be used on a job. Accordingly, those who have acquired more skills through attending school for longer periods are rewarded by society for their perseverance and higher skill levels through higher income.

The research hypothesis tells us what we should do empirically if a theory is useful, and if it is not. The human capital theory of education suggests that people who have more schooling should also earn higher incomes. We can look at real-world data and determine that this is wrong—and hence the theory is not helpful. Or we can look at real-world data and see that (in one sample) the human capital theory provides an explanation. Based on the results of our investigation, we may revise the theory or suggest an alternative theory.

Where do statistics fit in?

In the *exploratory* phase of research we collect data and seek to find patterns and make empirical generalizations. While there are many methods for doing this, descriptive statistics are often a very helpful tool. Often our exploratory data are not collected using probability samples or even highly standardized measurement tools—so inferential statistics are not very relevant, and we cannot apply hypothesis testing and prediction.

Once we have developed a theory and deduced a research hypothesis, we enter the *confirmatory* phase of research. Using research methodology tools, we make our theoretical concepts operational and develop measurable variables. Using research design and sampling, we collect systematic data, usually using probability or random assignment methods. With our data in hand, we can formally test hypotheses leading to verification of the theory and use statistical models to make predictions.

CHECK QUIZ 1.3

1. Social scientists use statistics to
   a. prove that research theories are true.
   b. manage and analyze data.
   c. maintain the status quo in the research community.
   d. assure that no errors were made in research design.
2. In the research process, theory
   a. is unnecessary.
   b. is always fully developed before any data are gathered.
   c. is developed only after the data have been completely analyzed.
   d. attempts to explain the relationship between phenomena.
3. A hypothesis differs from a theory in that
   a. it is testable.
   b. it is true.
   c. it is more speculative.
   d. it is more abstract.

The Research Report

Most of your professors have at one time or another written scientific research papers that have been published in professional or peer-reviewed journals. Conducting research and disseminating findings is a fundamental job requirement of the professoriate in most research universities in the United States and around the world. As undergraduate students, you have probably heard the expression that professors must either "publish or perish." Outside of the university, anyone who works with data must also, in a sense, publish or perish. Your boss will want you to examine some data for patterns and make a report; you may collect data on customer preferences to be used for marketing; you may evaluate whether one program of medical treatment produces better results than another. Applied statisticians must also report their results.

While the details will vary with your problem and your context, reporting the results of research that uses statistics almost always follows a common format. In preparing answers to many of the chapter exercises in this book, you can also use the following template.

The Research Problem

Research reports generally begin with a clear statement of the aim, objective, or purpose of the study. Hypotheses or expected patterns are clearly stated in a testable format. Since precision is crucial in the scientific enterprise, concepts should also be defined. The research problem answers the question: what does the researcher wish to explain, predict, or investigate? A typical research problem could be stated as: "The aim of this research is to investigate whether divorced men are more likely to commit suicide than divorced women."

Methods

Next the researcher indicates the nature and source of data to be used in bringing evidence to bear on the research problem and related hypotheses. Analysts can either use primary data that they collect, for instance through surveys, or rely on secondary data (collected by others) that exist in many data archives. For example, the National Center for Health Statistics (NCHS; part of the U.S. Department of Health and Human Services) collects and disseminates several data sets that are freely available at the organization's website (www.cdc.gov/nchs). Students interested in studying issues of health, fertility, family growth, morbidity, and mortality will find rich data sets at the NCHS site. An especially useful data set is the National Health Interview Survey. Other data archives include the Integrated Public Use Microdata Series (IPUMS) project at the University of Minnesota (www.ipums.org). Here one finds census data for the United States from earliest times to the latest decennial census of population. Many social scientists rely on data from the Inter-University Consortium for Political and Social Research (ICPSR) at University of Michigan. Data from ICPSR are available to researchers (including students and faculty) at universities that are members of the consortium and pay their regular dues. The United States Bureau of the Census (under the U.S. Department of Commerce) is a great source of information on the U.S. population. To determine what is available, visit the Internet site at www.census.gov. Two notable sets of data collected by the census bureau are the Current Population Surveys and the American Community Surveys.

In the methods section of the research report, it is not enough merely to indicate the type and source of one's data. The researcher must clearly state which variables were used and how they were measured. Where applicable, distinctions should also be made between dependent and independent variables. It is advisable to specify the dependent variable first, followed by your primary independent variable(s). It is also in the methods section of the research paper that the analyst indicates which statistical technique was used in order to test hypotheses and answer the research problem. Material that you are learning in this class will find their greatest application in the methods section of a research paper. It is required that the researcher state not only the statistical technique(s) used to address the research problem, but the appropriateness of the technique(s). As you study and apply statistics using this book, we will make it clear under what circumstances certain techniques are appropriate or not appropriate. Data analysis and statistical computations done on the data will be mentioned in this section of the paper as well, along with any known shortcomings of selected statistical technique(s).

Results

After analyzing the data, findings are presented in this section of the research paper. Depending on the research problem and the statistical technique(s) used, the researcher presents his or her results with the aid of statistical tables and charts. In this section, one must indicate which hypotheses were supported by the data analysis and which ones were not. It is not enough to merely present results; they must be interpreted in an objective manner for the intended audience. It is a good idea to keep tables to a minimum and to write out the findings as clearly as possible.

Discussion

In this section of the report, the researcher summarizes the findings and then goes on to place them within the context of the research problem and any relevant theories. The analyst also indicates the implications of the findings for past and future research. It is good discipline for a researcher to point out unique contributions made by his or her study. In this regard, an important question that should be addressed in the discussion is the following: What do we know as a result of this research that we did not know beforehand? The analyst must also indicate whether and how the findings are consistent with past studies. In addition, he or she should note existing studies that contradict the findings, and try as much as possible within the context of the data analysis to suggest possible reasons for inconsistencies.

Depending on the nature of the research question, it is a good idea for a researcher to suggest policy implications of the findings. Questions such as the following could be addressed: How could findings be used to improve the human condition? How might findings be used to solve specific problems in society? What is currently going on that needs to be reexamined and perhaps stopped as a result of findings from this study? Finally, the researcher should close by acknowledging any limitations of his or her study and offer caution (if appropriate) about the extent to which results might be generalized.

At the end of the research process, the author must communicate the findings to a wider audience. This can be done in several ways. The author could submit the article to a scientific journal for peer review and publication. Some researchers may decide to give a lecture within the university, in the community, or at another university, organization, or agency. Still others might decide to issue a press release, although it is advisable to issue a press release only upon the article's acceptance by a journal. Greater credence is given to the results if an article has gone through the peer-review process, is about to be published, or has already been published.

BASIC ELEMENTS OF RESEARCH: UNITS OF ANALYSIS AND VARIABLES

Researchers using statistical method use empirical observations—or data—to describe, infer, test hypotheses, and predict. But what are data?

Data are information, but they are more than that. Data are the result of using systematic methods of research design, sampling, and measurement guided by research hypotheses derived from theories. Because the information is collected with a method and plan, it has a structure. Data are scores on variables, observed across units of analysis.

Units of Analysis

Units of analysis in research are the carriers of information. They are the entities that the researcher studies, and upon which he or she collects data and makes generalizations. In a study that seeks to explain why higher educational attainment leads to higher income, if a researcher collects data on persons, then individuals comprise units of investigation. Similarly, in a study to determine whether poverty rates in cities contribute to higher crime rates, cities would be units of observation or study.

An important principle in statistics is that conclusions based on one unit of observation may not be generalized to other units. For example, if a researcher uses cities to study crime rates and finds that cities with high levels of poverty tend to have higher crime rates, it would be a mistake to conclude on the basis of the same study that poor people are committing the crimes. This is an example of a logical error, called the *ecological fallacy*.

Social scientists may study different units of analysis. Individual persons are the most obvious unit of analysis, but research might focus on families, groups, church congregations, neighborhoods, cities, political units like states or nations, or whole societies. What is important is that the unit of analysis be consistent with the theory, and that we use systematic methods to identify the elements of the population of units (i.e., be able, in principle, to list each member of the population).

Generally, social scientists, as we discussed earlier, don't collect information from every member of a population of units of analysis. Instead of doing a census, they usually draw a sample. The *cases* or *observations* or *units of analysis* in data, then, are usually elements selected from some homogeneous and bounded population to which the researcher wants to generalize.

CHECK QUIZ 1.4

1. A "case" in statistics refers to
    a. something to keep your calculator and pencil in.
    b. one in a set of variables (e.g., gender).
    c. an entity for which we have some data.
    d. one in a set of values (e.g., male).
2. A case is also known as
    a. an observation.
    b. a variable.
    c. an association.
    d. a difficult situation.

Variables

Statisticians collect information or data on characteristics or attributes of units of analysis. These characteristics are referred to as ***variables.*** A variable is a characteristic or attribute that can take on different values (or vary) across cases. It changes from person to person, from county to county, from city to city, or from country to country. A variable can be distinguished from a ***constant***, which does not vary across observations. In the General Social Survey, for example, all cases share the same score on the attribute "resident in the United States," since the study is conducted only in the United States. Research aims at explaining change or variation. Since constants by definition do not change, they cannot be objects of study.

Every variable is made up of *levels* or *values.* The levels or values of a variable are the scores that cases might have on an attribute. The variable "sex" might have two levels, male and female. The variable "income" might have many values: $3,426.27 and $105,232.16 are two examples.

Kinds of variables are often distinguished by the role they play in research questions (dependent or independent), whether their scales of measurement are discrete or continuous, and their levels of measurement.

Dependent and Independent Variables

A ***dependent variable*** is the entity that a researcher wishes to explain or predict. For example, if an analyst wishes to explain why workers have different earnings, then earnings comprise the dependent variable. Other names that are used in statistics to describe dependent variables are outcome or response variables. An ***independent variable*** is a variable that is believed to explain or predict the variation in the dependent variable. Other names that are used to describe independent variables are explanatory variables, predictors or predictor variables, covariates, and prognostic factors.

To differentiate between independent and dependent variables, think again about factors that explain earnings differences among workers. Suppose a sociologist was to propose a hypothesis that educational attainment influences one's earnings. Educational attainment would be the independent variable and earnings would be the dependent variable.

It is important to realize that no variable is inherently dependent or independent. Whether a given variable is independent or dependent is dictated by the research objectives. In one research a variable might be independent, but in another the same variable could be dependent.

Discrete and Continuous Variables

Variables can be described as *discrete* or *continuous* depending on their unit of measurement. A variable is considered a ***discrete variable*** if its unit of measurement cannot be broken down or subdivided into finer or smaller units. For example, the variable "children ever born per woman" is discrete. Its unit of measurement is human beings or persons, and these exist only as whole numbers (integers). Children ever born could take on values ranging from 0 (no children), 1 child, 2 children, or 3 children to the highest possible number of children born, but this variable will never take on values such as 0.1 children or 2.5 children. Other examples of discrete variables are household size, number of living children, number of married couples, number of cars per garage, and so on.

Continuous variables are those whose numerical values can be broken down or subdivided into finer units almost indefinitely. Age qualifies as an example of a continuous variable in that it could be broken down into years, months, days, and beyond. Other examples are weight, height, time, income, educational attainment, and so on. Many other continuous variables are formed by taking ratios or rates. The homicide rate (homicides per 100,000 residents) is a continuous variable because it divides homicides by population, and the calculation could be carried out to any number of decimal points. One hallmark of continuous variables is that in addition to a researcher's ability to break them down into finer gradations, they can assume decimals. In practice, many researchers end up rounding their values to one or two decimal places.

Levels of Measurement

The way in which we attach scores to attributes when we measure phenomena is very important. The ***level of measurement*** of a variable describes the kind and amount of information it contains; it also affects what kinds of operations we can perform on the variable, and the types of statistics that are appropriate.

There are four main levels or types of measurement: nominal, ordinal, interval, and ratio. The level or type of measurement to use on a variable is determined by noting the

CHECK QUIZ 1.5

1. A variable contains
   a. cases.
   b. values.
   c. statistics.
   d. observations.
2. Which of the following could *not* be a value?
   a. twelve (12)
   b. tall
   c. hairy
   d. often
   e. All of these could be values.
3. A researcher asks: "Are men or women more likely to become depressed?" The dependent variable is __________, and the independent variable is __________.
   a. women; men
   b. depressed; not depressed
   c. sex; depression
   d. depression; sex
4. The number of points that Kobe Bryant scores in a particular game is a(n) ___________ variable; Kobe's scoring average for the whole season is a ___________ variable.
   a. independent; dependent
   b. discrete; continuous
   c. outcome; predictor
   d. continuous; discrete
5. Suppose I am studying why some men drop out of high school and others don't. Dropping out of high school is a(n) ___________, and sex is a(n) ____________.
   a. constant; variable
   b. variable; constant
   c. independent variable; dependent variable
   d. dependent variable; independent variable

presence or absence of four characteristics: distinctiveness, ordering in magnitude, equal intervals, and an absolute or natural zero. Table 1.1 summarizes how levels of measurement distinguish types of variables.

Sometimes the levels of measurement are referred to in somewhat different ways. A variable that distinguishes kind but not amount is often referred to as "qualitative" or "categorical." Nominal and ordinal variables are qualitative—we can tell whether two cases

TABLE .1 Levels of Measurement and Their Characteristics

Characteristic	Nominal	Ordinal	Interval	Ratio
Distinctiveness	Present	Present	Present	Present
Ordering in magnitude	Absent	Present	Present	Present
Equal intervals	Absent	Absent	Present	Present
Absolute zero	Absent	Absent	Absent	Present

are of the same type but not the amount of the difference between them. A variable that distinguishes the amount of difference between cases is sometimes called "quantitative." Interval and ratio variables are quantitative.

Nominal variables have only the characteristic of distinctiveness. A nominal variable measures identity or category. In this level of measurement, numbers serve merely as labels or names to identify items, classes, or categories of the concept being measured. For example, if we are studying religion, we may sort our sample of people into categories and assign a number to each category, with 1 = Protestant, 2 = Catholic, 3 = Muslim, 4 = other, and 5 = none. These numbers serve to identify people by their religion. Apart from this, they mean nothing. For example, although the Catholic score is given as 2 and the Protestant score is 1, it does not make sense to assume that Catholics have more religion than Protestants, that there are twice as many of them, or any other conclusion based on the numbers 1 and 2.

A nominal variable that has only two categories (e.g., true/false, present/absent, yes/no, agree/disagree) is a *binary* (i.e., two-value) variable or ***nominal dichotomy variable type***. Usually one category is given the numerical value of 1, and the other category is given the value of 0. When scores are assigned this way, the variable is often called a *dummy variable.* A nominal variable that has more than two categories may be called a ***nominal polyotomy variable type*** (i.e., many types). Religion, for example, is a nominal polyotomy.

Ordinal variables indicate not only distinctiveness, but also ordering in magnitude. In such a case, larger numbers represent more of the concept or phenomenon being measured than smaller numbers. Numbers reflect the rank order of the concept. For example, suppose we want to measure the concept social class; we might divide our sample into 1 = lower, 2 = middle, and 3 = upper. It is clear that there is an ordering in magnitude, with those given 3 being of a higher class than those with 2. At the same time, the distance (or amounts) between classes are not necessarily equal. In this example, not only have we divided a concept (social class) into categories, but we have also shown that there are degrees of class.

Most ordinal variables used in social research are ***grouped ordinal variable type***. There are a limited number of ranks, and many cases may have the same rank on a grouped ordinal variable. In survey research, for example, respondents are asked whether they "strongly disagree," "disagree," "feel neutral," "agree," or "strongly agree" with a statement. We can order the respondents from low to high agreement, so the variable is ordinal. But many

CHECK QUIZ 1.6

1. The values “left-handed,” “right-handed,” and “ambidextrous” form
   a. a nominal dichotomy.
   b. a nominal polyotomy.
   c. a grouped ordinal variable.
   d. a full rank-order scale.
2. The values “Green Party,” “Democratic Party,” “Republican Party,” and “Chaos Party” form
   a. a nominal dichotomy.
   b. a nominal polyotomy.
   c. a grouped ordinal variable.
   d. a full rank-order scale.
3. The values “tallest in class,” “second-tallest in class,” and “shortest in class” suggest the use of
   a. a nominal dichotomy.
   b. a nominal polyotomy.
   c. a grouped ordinal variable.
   d. a full rank-order scale.
4. If a variable contains the values “failing course,” “just barely passing course,” and “doing really well in course,” it is probably
   a. a nominal dichotomy.
   b. a nominal polyotomy.
   c. a grouped ordinal variable.
   d. a full rank-order scale.
5. The original data had a variable called “Party” that contained the values “Green Party,” “Democratic Party,” “Republican Party,” and “Chaos Party.” We created three new variables out of the original variable. Now we have a “Green” variable (1 = Green Party, 0 = other), a “Democrat” variable (1 = Democrat, 0 = other), and so on. Forming variables in this way is known as
   a. stupid coding.
   b. ridiculous coding.
   c. dummy coding.
   d. Morse coding.
6. If you ask “How many years of school have you completed?” on a survey, this variable will be
   a. an interval/ratio variable.
   b. a full rank-order ordinal variable.
   c. a nominal dichotomy.
   d. a nominal polyotomy.

respondents can give the same answer, forming a group. *Full rank-order ordinal variables* are less common in social research. On a ***full rank-order ordinal variable***, each case has a distinct rank. The order of finish in a race (first, second, third, . . .) is a full rank-order ordinal variable because only one case can be at each level of the variable.

Interval variables have the characteristic of distinctiveness, ordering in magnitude, and, in addition, equal intervals. The variable age is an example of an interval variable. Imagine three people, one 16 years old, another 17 years old, and the third 18 years old. The three ages are different (and hence the distinctiveness quality exists). In addition, 17 is greater than 16, and 18 is greater than both 16 and 17. Thus there is ordering in magnitude with higher numbers denoting more age. What distinguishes this level of measurement from the rest is that in addition to the two aforementioned characteristics, there are equal intervals or equal distances between numbers. For instance, the distance (interval) between 16 and 17 is the same as that between 17 and 18.

Ratio variables have all the qualities previously stated (distinctiveness, ordering in magnitude, and equal intervals). Additionally, they have what is called an absolute or natural zero. This simply means that a zero represents a complete absence of the phenomenon or property in question. For example, in talking about income, zero income implies complete absence of money and hence income would qualify as a ratio level variable.

Take a minute, before moving on, to be sure you understand the distinctions among the levels of measurement that are used in statistical measures.

SUMMARY

The purposes of this chapter have been to clarify the role that statistics play in the process of scientific inquiry, and to review some basic concepts and terminology that you need to know before beginning to study statistics.

Knowledge of the logic and some of the technique of statistical analysis is simple cultural literacy in contemporary society. Statistics are used for a number of purposes: describing the variation and covariation of variables across cases, making inferences from observations in a sample to larger populations from which the sample was taken, testing hypotheses derived from theories of how the social world works, and making predictions based on models of the relationships among variables.

In this book we discuss statistics as a set of tools that support scientific inquiry. Scientific inquiry is primarily concerned with developing theory and validating theory by empirical observation. Statistics play roles both in creating theory and in validating theory by observation. The former (exploratory) role uses statistics as a tool to summarize data and find patterns so empirical generalizations can be made, which then beg theoretical explanation. The latter (confirmatory) role involves using theory to make predictions—hypotheses—about what we should observe in new samples if a theory is valid (or not), and then using statistics to determine the confidence we have in asserting that observations are (or are not) consistent with theoretical expectations.

In this chapter, we've also provided an outline of how statistical information is used in writing a research report. Statistics are not an end in themselves, and the research report is a good example of how they are actually used in practice in social-scientific research.

Statistics are operations performed on data. Data are information describing scores on variables that are attributes of units of observation. *Cases*, *units*, and *elements* are terms used to identify the units of observation in data. Social scientists may study many different kinds of units.

Each unit of observation is measured, and then assigned scores to describe its attributes. Measured attributes that may differ from one unit of observation to another are called variables. Variables have levels or scores that indicate the possible states of each attribute. It is often useful to distinguish between independent and dependent variables, discrete and continuous variables, and qualitative and quantitative variables.

Each variable has a level of measurement or scale. The scale used to assign numbers to the levels of a variable is important because different scales contain different kinds of information and allow different kinds of statistical procedures. We discussed nominal, ordinal, interval, and ratio scales for the levels of variables.

We've finished with the preliminaries, so let's start doing some statistics!

KEY TERMS

dependent and independent variables

descriptive statistics

discrete versus continuous variables

full rank-order ordinal variable type

grouped ordinal variable type

hypothesis testing

inferential statistics

interval variable type

levels of measurement or scale

nominal dichotomy variable type

nominal polyotomy variable type

prediction

ratio variable type

sample and population

statistic and parameter

variable versus constant

CHECK QUIZ ANSWERS

Quiz 1.1 Answers: 1. (a); 2. (d); 3. (a)
Quiz 1.2 Answers: 1. (b); 2. (d); 3. (c); 4. (a)

Quiz 1.3 Answers: 1. (b); 2. (d); 3. (a)
Quiz 1.4 Answers: 1. (c); 2. (a)
Quiz 1.5 Answers: 1. (b); 2. (e); 3. (d); 4. (b); 5. (b)
Quiz 1.6 Answers: 1. (b); 2. (b); 3. (d); 4. (c); 5. (c); 6. (a)

EXERCISES

1. What are descriptive statistics used for? What are inferential statistics used for? Provide an example of each.
2. For each of the following scenarios, determine whether the use of statistics being described is descriptive or inferential.
   a. The average age of students in a class is found to be 19.4 years.
   b. A poll of public opinion reports that 38 percent of the population supports ballot proposition B, with a margin of error of 3 percent.
   c. The annual family incomes of persons applying for disability income in San Rafael during 2003 ranged from $15,437 to $234,550.
   d. Based on a sampling of students enrolling in 15 classes taught in a large lecture hall, the registrar estimates that 38 percent of all students enrolled in classes in the lecture hall are "extremely dissatisfied" with the acoustics.
3. Here are a number of variables that are used in the General Social Survey. For each variable, indicate the level of measurement (nominal, ordinal, or interval/ratio) and whether the variable is continuous or discrete.

Variable	Level of Measurement	Continuous or Discrete
What is your sex? (Male or female)		
How many siblings (i.e., brothers and sisters) do you have?		
What social class would you say your family is? (Upper, middle, working, or lower)		
Do you believe in life after death? (Yes, not sure, no)		
What is your current grade point average (GPA)?		
What is your mother's occupation?		
How many years of schooling did your father complete?		
The death penalty for murder should be used in all states. (Strongly agree, agree, neutral, disagree, strongly disagree)		
How many times per year do you attend religious services?		

4. For each of the variables listed, state the level of measurement and whether the variable is continuous or discrete.

Variable	Level of Measurement	Continuous or Discrete
Student's age, measured in years		
Proportion of the students in classes at the university who are women		
Student's class (e.g., first-year, senior)		
Self-identified national origin (e.g., United States, Cambodia, Senegal)		
Father's education (less than high school, high school, some college, completed college, or graduate work)		
Support for the war in Iraq (strongly support, support, neutral, oppose, strongly oppose)		
Grade point average, measured to two decimal places (e.g., 3.45)		
Member of fraternity or sorority (yes, no)		
Number of times per week that you eat at an on-campus dining facility		
Number of persons in your immediate family		

5. The following are some short descriptions of research studies. For each, identify the variables that are being used, describe the level of measurement of each, and determine whether the variable is being used as an independent or a dependent variable:
   a. A student was recently arrested for hacking into the database used to conduct elections for officers of the senior class. The campus newspaper has asked you to do a study to determine how other students feel about this—that is, how many students feel that this is a very serious crime, not a serious crime, or not really a crime at all. The editor thinks that computer science students and business students are less likely than other students to view the hacking as a serious crime.
   b. The campus computing service personnel are interested in how aware students are of the services they offer. They ask a sample of 500 students whether they are aware (yes, maybe, no) of each of 10 different services, and count up how many are chosen. To try to target future advertising, they also ask respondents to report their sex, age, and year in school, and whether they own a computer.
   c. A political economist has derived the hypothesis that "the greater the level of income inequality, the greater the political instability." To test this hypothesis, she collects data on the percentage of all income that is controlled by the richest 5 percent of families in each of 100 nations in the year 2000. The researcher measures

the concept of "political instability" as a count of the number of changes in government between 2001 and 2004.

d. A researcher thinks that police-community relations are affected by city size and racial heterogeneity. The researcher measures the quality of police-community relations by doing a survey in each of 30 cities, asking the question: "How would you rate your local police? (Excellent, good, fair, poor)." City size is measured by the number of people in the city as reported by the census. Racial heterogeneity is measured by the ratio of nonwhite persons to white persons.

6. The following are some short descriptions of research studies. For each, identify the variables that are being used, describe the level of measurement of each variable, and determine whether each variable is being used as an independent or a dependent variable:

   a. A researcher is interested in understanding why some students earn higher grades than others. One possible explanation is that performance in college is a continuation of earlier academic performance. The researcher collects information on a sample of students, measuring their current GPAs and their high school GPAs.

   b. Some communities have higher rates of crime than others. Our researcher thinks that high average income and high average education of the families in communities may decrease the rates of some crimes (e.g., family violence), but increase rates of other crimes (simple theft).

   c. As organizations become larger, they need to become more vertically differentiated (i.e., have more levels of management). A researcher collects information on the total employment of universities and the number of levels of administration in order to test this hypothesis.

   d. Does ethnic identity become less the longer a family has been in the United States? A researcher asks a series of questions designed to measure ethnic consciousness in people of Japanese origin whose families have been resident in the United States for one, two, three, four, or more generations.