Introduction 1

In this chapter, you can learn

- what you can find out about data analysis from this book,
- why data analysis is important,
- how much math and computer knowledge you will need to use this book,
- some concepts from research methods that are essential for data analysis, including
    - cases and units of analysis;
    - variables, attributes, and levels of measurement;
    - reliability and validity;
    - independent, dependent, and control variables;
    - requirements for demonstrating causality; and
    - probability and nonprobability sampling.

Welcome to *Answering Questions With Statistics.* This book has been written to help you learn data analysis: how to do it yourself and how to understand what others have done. Since this is probably your first comprehensive introduction to data analysis, the material is presented in a simple, straightforward way. While there is much to learn, most is easy to understand. To be successful at mastering data analysis, you don't need to be a math whiz, but you do need to be disciplined. You need to read the chapters carefully, follow the examples in the chapters, and do the practice problems. By doing that, when you complete this book, you will understand and be able to do the statistical procedures most commonly used by social scientists. So, let's begin.

About Learning Statistics

Statistics Are Tools for Answering Questions

Statistics are just tools. They help us answer questions and make decisions. While you can admire the beauty of statistics and the elegance of the math behind them, most people who use statistics just want to find something out. Statistics are a means to an end, not an end in themselves. That is the approach this book will take.

RESEARCH TIP

What's the Difference Between Statistics and Data Analysis?

Statistics focuses on the properties of the statistic itself. It explains why the statistic works. The statistic is an end in itself. Data analysis takes a more applied approach. It emphasizes how to correctly use a statistic to answer a question. The statistic is a means to an end. This textbook is all about data analysis.

Almost every example in this book and almost every practice problem flow directly from one central question: Have young adults changed in 30 years? This question can be taken two ways: "How do the young adults of today compare to the young adults of 30 years ago?" and "How have the attitudes and behaviors of those young adults of 30 years ago changed now that they are middle-aged?" These questions are relevant to anyone of any age who wonders how America is changing and how they themselves have changed or will change as they grow older.

The companion data file to this book, which can be downloaded from the book's website, has information on 1,258 persons who were respondents to the General Social Survey (GSS). The GSS is a national sample survey of adult Americans funded by the National Science Foundation and currently carried out every other year by the National Opinion Research Center. Our 1,258 persons represent four distinct groups: *1980 young adults, 1980 middle-age adults, 2010 young adults, and 2010 middle-age adults.* As the names indicate, the first two groups come from the 1980 GSS, the last two from the 2010 GSS. The two young adult groups were in their 20s when they took part in the GSS; the two middle-age adult groups were in their 50s. The four groups come from three different generations: The 1980 middle-age adults are from America's World War I generation, the 1980 young adults and the 2010 middle-age adults are from the same baby boom generation at two different points in time, and the 2010 young adults are from Generation Y. Table 1.1 shows the actual years of birth for the groups and some alternate names that will sometimes be used for the groups.

With data on these four groups, we can ask and answer some interesting questions: What was each of the four groups like? How do 2010 young adults compare to 1980 young adults? How does the generation gap between young and middle-age adults today compare to the generation gap of 30 years ago? And how have the 2010 middle-age adults changed from the 1980 young adults they were? (The fiftysomethings who answered the GSS in 2010 and the twentysomethings who

RESEARCH TIP

More About the General Social Survey

For more information about the General Social Survey, visit http://www.norc.org/projects/General+Social+Survey.htm.

The complete citation to the data file from which our 1980 and 2010 data came is as follows: Tom W. Smith, Peter V. Marsden, Michael Hout, and Jibum Kim: *General Social Surveys, 1972–2010* [machine-readable data file]. Principal Investigator, Tom W. Smith; Co-Principal Investigators, Peter V. Marsden and Michael Hout, NORC ed. Chicago: National Opinion Research Center, producer, 2005; Storrs, CT: The Roper Center for Public Opinion Research, University of Connecticut, distributor. 1 data file (55,087 logical records) and 1 codebook (3,610 pp).

Table 1.1 The 1,258 Persons We Will Study Are Divided Into Four Groups

Group	Year of Birth	GSS Year	Age at the Time of the GSS	Number of Persons in Data Set
1980 young adults (1980 twentysomethings)	1951 to 1960	1980	20 to 29	327
1980 middle-age adults (1980 fiftysomethings)	1921 to 1930	1980	50 to 59	212
2010 young adults (2010 twentysomethings)	1981 to 1990	2010	20 to 29	341
2010 middle-age adults (2010 fiftysomethings)	1951 to 1960	2010	50 to 59	378

answered the GSS in 1980 are not the same individuals since the GSS takes a new random sample of the population each time it is conducted, but they come from the same baby boom generation at two points in time: young adulthood and middle age.)

The data analysis examples in the chapters will show you what the 1980 young adults were like. To find out what the 2010 young and middle-age adults were like and how much change has occurred in 30 years, you will have to do the practice problems at the end of each chapter. Practice problems are an essential part of learning data analysis. Inevitably, you learn by doing.

What Can You Expect to Learn About Data Analysis?

As you find out about young adults and how they have changed, you will be learning

- how to understand statistical information when it is presented to you and
- how to produce meaningful statistical information.

This book's first purpose is to make you an informed consumer of statistics. Almost everywhere you look, you see statistical information. It is in newspapers and on the evening news. It is certainly in textbooks and in professional journals. You will find information being communicated in statistical form in reports done by you or for you in your job. We have all heard how easy it is to lie with statistics, but that depends on how much the person receiving the information knows about statistics. A person who knows nothing about statistics can be easily fooled, but to fool a person who knows her way around data analysis is not so easy.

This book's second purpose is to make you a competent producer of statistical information. To do this, you will be introduced to and use one of today's most commonly used statistical software packages—*IBM®SPSS® Statistics.* It is used in colleges and graduate schools, in government offices, in welfare agencies and police departments, and in private businesses. ("SPSS" is the shorthand way that we will refer to *IBM®SPSS® Statistics.*) Knowing how to use SPSS is a concrete skill that employers value. Once you have mastered the SPSS techniques described in this text, you should indicate on your resume that you have knowledge of SPSS. It is that valuable!

Why Is Data Analysis Important?

You are probably taking a course in data analysis because your major department requires it. That should tell you data analysis is pretty important. Many licensing exams in social service professions test for the material taught in data analysis, and many graduate and professional programs build on the skills taught in data analysis.

But why do your major departments want you to take data analysis? Why do licensing exams test for data analysis knowledge, and why do graduate and professional programs want you to learn even more about data analysis? Some of the reasons are that understanding data analysis makes you

- a better citizen,
- a better professional,
- a better administrator, and
- a better policy maker.

As a citizen, you have a responsibility to understand the issues of the day, to comprehend the extent of social problems and the effectiveness of current programs, and to choose between the proposals of competing political candidates. A democracy depends on an informed electorate. You cannot consider yourself informed in the 21st century without a working grasp of statistics.

As a professional, you have an ethical responsibility to provide the best available service to your clients—whether those clients are recipients of welfare programs, students in a classroom, or citizens seeking safety under police protection. Hopefully you are a concerned, caring individual, but that is not what will make you a professional. Knowing the effectiveness of alternative treatment programs and administering the best available treatment is what makes you professional. You cannot do that without knowledge of statistics.

As a supervisor of other professionals—and many of you will rise to managerial positions—you will be called upon to make decisions about your agency's staffing, your office's programs, and your department's effectiveness. Without basic data analysis skills, you will lack the tools necessary to make the best decisions.

Important policy decisions in our businesses, communities, and governments are not made solely on statistical evidence. But without a grasp of statistical techniques, your chances of getting a seat at the table where important decisions are made are slim, and your chances of holding that seat, should you happen to get it, even slimmer.

How Much Math or Computer Knowledge Do You Need?

As far as this text is concerned, you knew enough math when you graduated from high school and probably from grade school. This is not a math text. While you will see the formulas for some of the simpler statistics and occasionally use those formulas to calculate statistical values, you will usually depend on the computer to calculate those values.

But what if you have forgotten most of the math you knew when you graduated high school? Take heart—much of it comes back very quickly. Having said that, if you are seriously concerned about your math skills, here are some suggestions:

- If you have not yet taken and passed a college-level math course, any college-level math course, do that before taking data analysis. Any math course will remind you of those math principles you knew when you graduated from high school.
- Find a math review book for students preparing for college entrance exams such as ACT, SAT, or even a college entrance basic skills test. Ignore the geometry, ignore the trigonometry and any calculus, and ignore any advanced algebra. Just review the arithmetic and the simplest algebra.

Don't worry! You almost certainly know enough math.

But do you need to have taken a course in probability and statistics before taking data analysis? The more you know about probability theory and formal statistics, the better you will understand how the statistics described in this book work and why you can have confidence in them. But do you need a probability and statistics course to understand and master the data analysis techniques described here? No.

So, if you have taken a probability and statistics course, will you learn anything new here? Yes. Because this book does not go into the theory behind particular statistics, it has time to introduce more statistics than are usually covered in probability and statistics courses. More important, you will see the statistics being used in context—specifically in a social science context. You will see in what situations social scientists choose which statistical tools and how social scientists move from statistical results to decisions.

Since SPSS is a computer software program, do you need to know a lot about computers or programming? Absolutely not! SPSS is not a programming language like Java, C++, or Pascal. Those are computer languages in which users write commands observing the language's specific rules of syntax. SPSS is a statistical application. It consists largely of pulling down menus, making selections, pointing and clicking. Although most versions of SPSS do provide users with the option of writing syntax commands, which can dramatically save time when doing certain complex or often repeated procedures, we will stick with just pulling down menus, making selections, and pointing and clicking.

The mechanics of using SPSS are few in number and simple to learn. After just a few chapters, getting SPSS to do what you want will be no problem. The bigger tasks will be deciding what you want SPSS to do and drawing conclusions from the output SPSS provides you.

CONCEPT CHECK

Without looking back, can you answer the following questions:

- Who are the four groups we will be comparing in the chapter examples and practice problems?
- In what roles will you be better because you understand data analysis?

If not, go back and review before reading on.

Looking Ahead

The rest of this chapter reviews some important research methods concepts. Chapter 2 then walks you through a simple SPSS session using an already existing data set, and it shows you how to create an SPSS data set on your own. Once you understand these things, you are ready to begin answering questions with statistics. Starting with Chapter 3, chapters start off with questions about 1980 young adults. Each chapter presents the data analysis techniques needed to answer those questions. The practice problems at the end of each chapter then give you a chance to use those techniques to see how young adults have changed in 30 years.

Essential Concepts From Research Methods

Before you can use data analysis to answer questions, you need to understand some things about how data are collected. Data collection techniques are often referred to as research methods. Data analysis and research methods go hand in hand. Which statistical techniques you can use depend on how the data were gathered. In turn, decisions about how to gather data depend on which statistical techniques you want to use. It is the classic chicken-and-egg question about what to study first: research methods or data analysis. That is why some schools have students learn methods first and others data analysis first.

The remainder of this chapter describes some concepts from research methods that are particularly important for doing data analysis. Whether this is a review for you of material you covered before or a preview of what you may be studying in more depth in the future, read carefully. The terms and ideas described in this chapter are used frequently in the chapters ahead.

Cases

Science proceeds by means of comparison. Social scientists rarely study just one person or one family or one community. They study many individuals (or families or communities), looking for ways in which they are similar and ways in which they differ and then seeing if some of those differences are related. And, if they are related, they try to understand why.

In survey research, each individual who completes a questionnaire represents a separate **case.** In a study of U.S. states, each of the 50 states would represent a case, and the researcher's data set could be described as consisting of 50 cases.

Depending on the researcher and the type of study being done, cases may also be referred to as subjects, participants, or observations.

Statistical analysis can reveal patterns in the characteristics of cases—patterns that the unaided researcher might miss. The greater the number of cases being studied, the harder it is to detect subtle or complex patterns in the data, and social scientists often analyze data sets containing hundreds or thousands of cases, making statistical analysis essential.

The data analyst needs to know what the cases in a data set represent. In the majority of social science data sets, the cases represent individual persons. But in some data sets, the cases represent something else. The cases might represent families, occupations, unions, nations, or many other possibilities. The **unit of analysis** is simply a way of referring to what the cases in a study are. Knowing what the cases in a data set represent is important because the analyst is justified in drawing conclusions only about the units of analysis for which she has data. For example, if we have data on counties, we can legitimately talk about the characteristics of counties and how certain county characteristics are related to other county characteristics. If we were to talk about the characteristics of individuals and how those individual characteristics are related when all we have are data describing counties, we would be exceeding the limits of our data and committing what is termed an **ecological fallacy,** which is the assumption that what is true about groups must inevitably also be true of the members of those groups.

Data sets generally do not mix together different units of analysis. A data set of 1,500 cases might consist of data on 1,500 individual persons or 1,500 different families, but it would not consist of data on mostly individuals but with a few families thrown in. Whenever you are using a data set for the first time, be sure to find out how many cases there are and what the cases represent.

Variables

Scientists are interested in the differences between cases. For example, families differ in income and in type of residence. Universities differ in the selectivity of their admission standards and whether they are public or private. Any dimension on which the cases in a study differ is known as a **variable.**

As you can imagine, cases can differ in many ways. Individuals, for example, differ physically; they also differ in attitudes and values, the frequency with which they do certain things, in their religious and political affiliations, and so on. No study records all the variables on which cases differ.

A researcher decides which variables to study based on a theoretical perspective, a review of the literature, logical thinking, a hunch, or preliminary unstructured observation.

Deciding which variables to include in a study and which to exclude is tough. Every variable included will require more time, energy, and money devoted to data collection. Every variable excluded will be a possible answer to a question that will go unexplored.

Data sets differ in the number of variables they contain. A data set could consist of just a single variable but usually contains more than that. Data sets based on answers to large surveys often include hundreds of variables.

Theoretical Definition

Variables are usually referred to by a brief name such as *gender, social class,* or *occupation,* but that name is really just an abbreviation. A researcher needs to know, and anyone reading the researcher's results needs to know, what the researcher means by that name. When the researcher says *gender,* what does he mean by that term? Or when she says *social class?* Or *occupation?* Each variable, besides having a name, needs to have a **theoretical definition.** Sometimes called a nominal definition, a theoretical definition is like the definition you would find in a dictionary. It explains in abstract terms what the researcher means by a certain variable name.

Two researchers might theoretically define the same variable name in different ways. There is absolutely nothing wrong with that, although it will complicate communication between the researchers. The important thing is that each researcher and her audience know what is being meant when that researcher uses that variable name.

Operational Definition

A theoretical definition clarifies for both the researcher and his audience what is meant by a particular variable name. But using that variable in an empirical study requires more. A researcher must know and be able to explain to others precisely how that variable is going to be measured. The description of the procedure by which a variable is going to be measured is known as the **operational definition.**

A good operational definition should indicate how the measurement will be done. Will it be based on observing the case or by asking a question of the case or, perhaps, by examining something produced by the case such as a journal entry? If you are going to measure a variable by asking a question, what specifically will that question be and, very importantly, how will the person indicate his answer? Will he give an unstructured answer or choose from designated categories and, if the latter, what are those categories? Will the variable be measured by asking just one question or will multiple questions be needed, and if multiple questions are required, how will the several answers be combined into an overall score?

Even if two researchers agree on the theoretical definition of a variable, they might operationally define it differently. That is their right, although it will again complicate communication between the researchers and make a straightforward comparison of their results difficult. The important thing is that each researcher and his audience know what is being measured when that researcher uses that variable name.

Within a single data set, researchers should not change the operational definition they are using for a variable. You cannot measure readiness for college by administering the ACT exam to half your subjects and the SAT exam to the other and treating the scores as if they all came from the same operational definition.

Attributes

Part of the operational definition is a clear specification of the categories or **attributes** that make up the variable. These are the possible scores or values cases may receive on the variable. For example, the variable gender would probably have the categories male and female. Social class might be operationally defined as consisting of the categories lower class, middle class, and upper class. The variable number of children might be operationally defined as consisting of the categories 0, 1, 2, 3, 4, 5, and on up to the highest integer needed. Normally, the attributes of a variable are constructed so that every case will find one and only one category that fits. (Having cases fit into one and only one category is sometimes accomplished by instructions on survey questions or directions to observers that state to use one and only one answer category.)

Once you know the attributes that make up a variable, all the other information in the operational definition about whether to ask questions or observe, what question or questions to ask, and how to combine answers to several questions can be thought of as the procedure to follow to figure out which is the correct attribute for each case. It is very important that the same procedure be followed for each case. Although the result of applying the procedure to different cases may be different—that is, some cases end up in one category while other cases end up in other categories—the procedure to determine the appropriate category needs to be applied consistently. If the procedure is done one way for some cases but another way for other cases, then the procedure itself rather than the actual differences in the cases may be determining into which category the case is placed.

As noted earlier, researchers may operationally define the same variable differently. That extends to differing on the number of attributes that make up the variable. For example, take the variable age. One researcher may operationally define the variable age as the number of full days a person has lived since birth. That would result in a great many attributes since a centenarian has lived approximately 36,525 days! Another researcher might operationalize the same variable as age at last birthday. That would result in far fewer attributes. The researcher who uses age at last birthday but defines the attributes in terms of categories (0 to 9, 10 to 19, 20 to 29, 30 to 39, 40 to 49, and so on as needed) would have still fewer attributes. Still another researcher might operationalize age as young, middle aged, and old. One way of operationalizing age is not necessarily better than another. However, how a variable is operationalized will affect which statistics can be used in its analysis.

Variables differ in the number of attributes that make them up. In the simplest situation, a variable has just two attributes, for example, female and male, yes and no, agree and disagree. A variable with just two attributes is known as a **dichotomy.** Dichotomies sometimes receive special treatment in data analysis. Of course, a variable could not have just one attribute. Variables are dimensions on which cases differ. If there is but one attribute, then there is no possibility for cases to differ. A dimension on which cases do not differ either because there is no alternative category or because all the cases fall into just one category is referred to as a **constant.**

At the other extreme, variables may have a very large, even an infinite, number of attributes. If the variable is population size and the cases are nations and the attributes are the set of positive integers (1, 2, 3, 4, 5, and on up as high as needed), we would have more than 1,300,000,000 attributes since China has a population in excess of 1.3 billion persons. Now, you might say, that's true, but many of those attributes won't be used. Since there are only about 200 nations in the world, no more than 200 attributes, and possibly less if any countries have identical populations, will have any cases in them. However, the list of attributes that make up a variable does not depend on which attributes are actually used in a data set. An unused attribute is still an attribute for that variable.

Researchers distinguish between discrete and continuous variables. A **discrete variable** has a finite and usually small number of attributes, whereas a **continuous variable** has a large, theoretically infinite, number of attributes. The distinction is an important one because discrete and continuous variables sometimes require different statistical techniques. Continuous variables are always based on a numeric scale of measurement, but not all numeric scales are continuous variables. Continuous variables have a potentially infinite number of possible values between any two attributes. For example, the percentage of a population that is female could have a value (or attribute) of 55%. It could also have a value of 56%. Between those two values, however, there are an infinite number of possible values. The percentage female could be 55.5%, 55.25%, 55.125%, and so on. By comparison, consider the variable number of children. Both 3 and 4 are legitimate values. But there are no values between 3 and 4 that are possible. No one has 3.5 children. Thus, number of children is a discrete variable. Had a researcher been comparing nations, however, and the variable been average number of children, fractional values become possibilities, and the variable would be considered continuous. Variables whose attributes are not numeric quantities such as political party affiliation (Democrat, Republican, other, none) or degree of agreement with a particular statement (strongly disagree, moderately disagree . . . moderately agree, strongly agree) are always considered discrete variables. Another term sometimes used for discrete variables is *categorical variables.*

Levels of Measurement

A fundamental concept for data analysis is **level of measurement.** Based on the properties possessed by a variable's attributes, the variable can be described as having a nominal, ordinal, interval, or ratio level of measurement. You set a variable's level of measurement, whether you realize it or not, when you operationally define the variable. Decisions made early in the research process such as the answer choices you are going to provide for a question in a survey, the recordkeeping procedures to be used by observers, or the classification systems to be used in analyzing artifacts determine a variable's level of measurement. Although the critical decisions are made early in the research process, the consequences of those decisions about level of measurement only show up when you are ready to do data analysis. You must know the level of measurement of every variable you use in your analysis because which statistical procedures are legitimate to use and which are not depend on the level of measurement of the variables involved.

In most research methods classes, four distinct levels of measurement are introduced: nominal, ordinal, interval, and ratio. For an introductory course in data analysis such as this one, however, the distinction between interval and ratio does not matter. None of the statistics introduced in this book require a ratio level of measurement. So, you will not need to distinguish between interval and ratio levels. Those last two levels are combined so that there are just three levels of measurement:

nominal, ordinal, and **interval/ratio.** (As you will soon see, SPSS gives the name *scale* to this combined interval/ratio level of measurement.)

Nominal is considered the lowest level of measurement, next comes ordinal, and interval/ratio is considered the highest. Nominal attributes have the least information in them, ordinal attributes have more information packed in them, and interval/ratio attributes have the most usable information. Correspondingly, nominal variables support the fewest legitimate statistical procedures, ordinal variables support more, and interval/ratio support the most.

In deciding a variable's level of measurement, you need to know the attributes that make up the variable. In deciding about level of measurement, the number of cases that fall into particular attributes makes no difference. Level of measurement can be determined before any of the data are actually collected.

Nominal

You can determine a variable's level of measurement by asking just three yes/no questions about the attributes that make up the variable. Figure 1.1 shows you the questions and what to do with the answers.

The first question is "Do the attributes cover all the possibilities without overlapping?" You want to create a set of attributes so that every case finds one and only one attribute that fits it. Sometimes this is accomplished with instructions like "choose one and only one of the following answers," "choose the one best answer," or "check the one category that comes closest to what you observed." In fact, unless told otherwise, you can assume respondents, observers, and examiners were told to choose one and only one attribute.

Figure 1.2 repeats the flowchart of questions in the previous figure but at each stopping point shows an example of a variable that would have traveled that path.

One stopping point is if the answer to the first question is no. The variable "personal income last year," as operationalized in Figure 1.2, does not even have a nominal level of measurement. The categories overlap. Someone with exactly $10,000 income could go into either of two categories; the same is true for someone with $30,000 or $50,000 income. Overlapping categories is a surprisingly common mistake in designing operational definitions, and it needs to be corrected before the data start to be collected. There is usually no good way to fix the problem after the fact. So be sure to carefully examine and pretest your research instruments before using them!

Now consider a variable that records the state in which a person was born. Its attributes appear in Figure 1.2. These attributes cover all the possibilities without overlapping. For this variable, the answer to the first question is yes.

If the answer to this first question is yes, then the variable's level of measurement is at least nominal. Mathematically, if you are comparing two cases on a nominal variable, you can state whether the two cases are equal (=) or not equal (≠). That may not seem like much, but it is a start and will allow you to generate frequency distributions.

Ordinal

The second question is "Can the attributes be put in a natural order from low to high?" Can the categories that make up the variable be arranged so they represent increasing amounts of what the variable measures? In practical terms, if you put each attribute on a card, mixed them up, and tossed them on

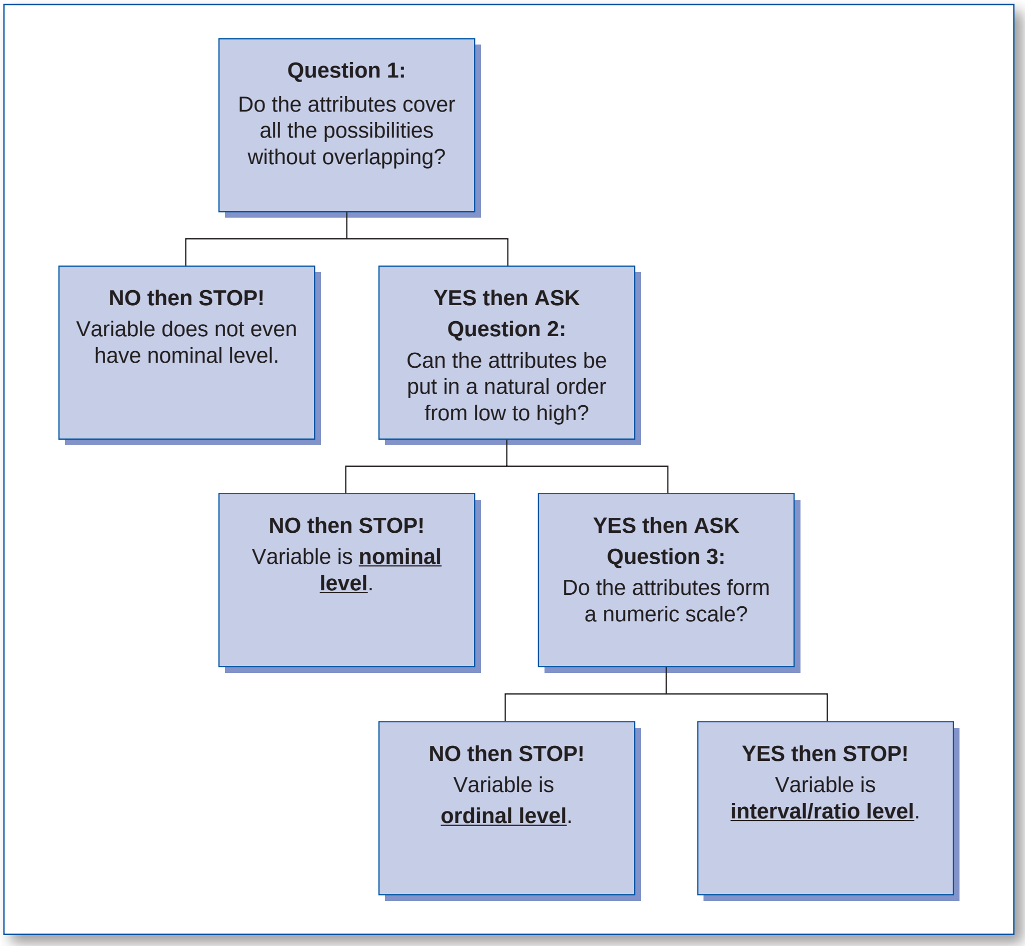

Question 1:
Do the attributes cover all the possibilities without overlapping?
NO then STOP!
Variable does not even have nominal level.
YES then ASK
Question 2:
Can the attributes be put in a natural order from low to high?
NO then STOP!
Variable is nominal level.
YES then ASK
Question 3:
Do the attributes form a numeric scale?
NO then STOP!
Variable is ordinal level.
YES then STOP!
Variable is interval/ratio level.

Figure 1.1 Determining a Variable's Level of Measurement

the floor, would a reasonable person be able to put them in order from low to high? If so, then the attributes have rank order. Remember that the order must be based on what the variable is trying to measure. Alphabetical order does not count nor does an order based on how many cases chose each attribute.

The attributes for state of birth do not possess a natural order. No matter how much you personally favor one state over another, the attributes are just different from one another. They lack a natural rank order. So the variable state of birth is a nominal-level variable.

Now consider a variable that records the extent of a person's agreement or disagreement with the statement "I am looking forward to learning more about data analysis." The attributes appear in Figure 1.2. Assume respondents were told to select the single answer that most accurately describes their attitude. For this variable, the set of attributes not only represents all the possibilities without overlapping but also can be put in order from least agreement (strongly disagree) to most agreement (strongly agree).

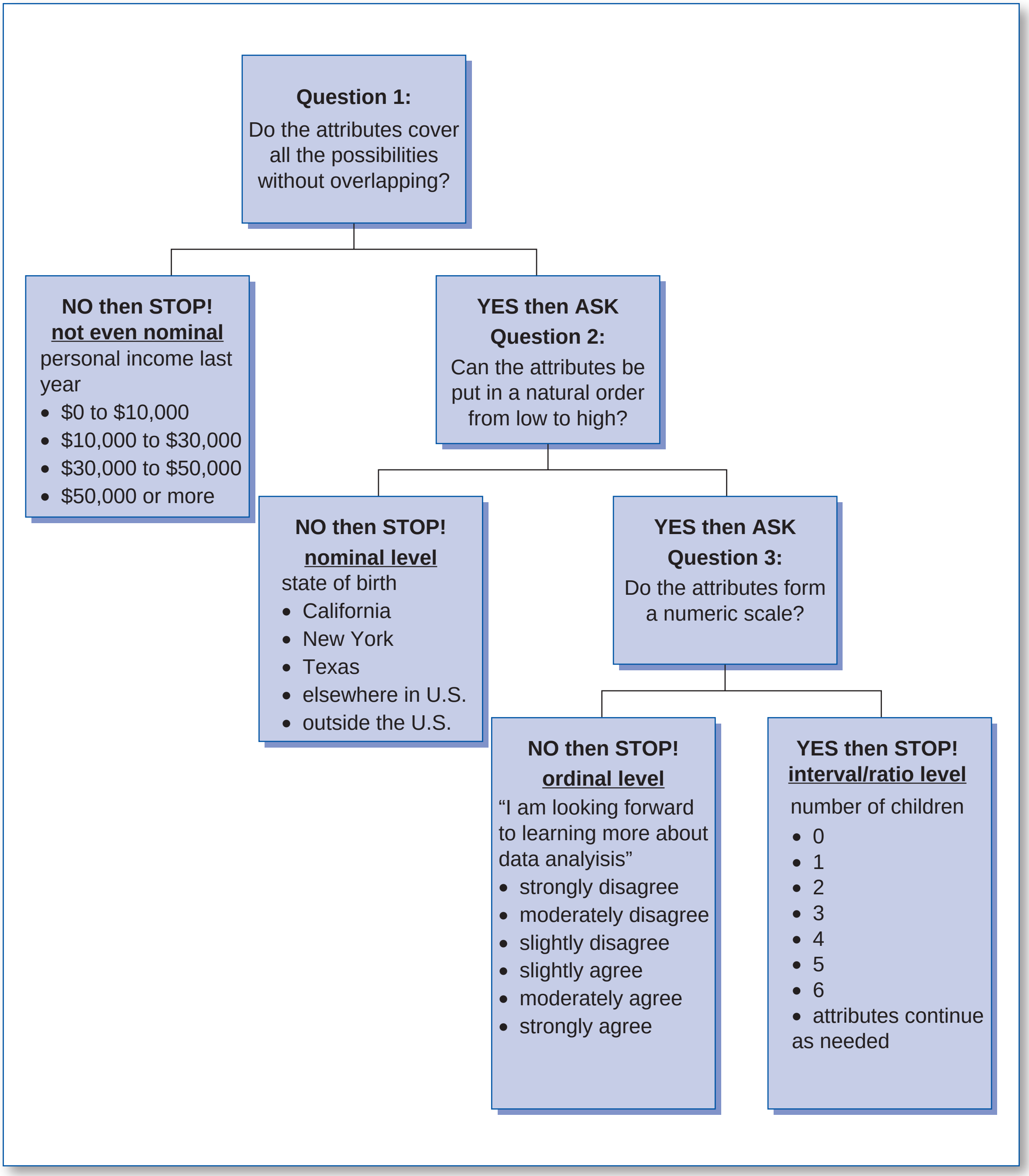


Figure 1.2 Determining a Variable's Level of Measurement With Examples

If the answer to the question about all possibilities without overlapping is yes and the answer to the question about a natural order is yes, then the variable is at least ordinal. When comparing two cases on an ordinal variable, not only can you state whether the two cases are equal or not equal, but if they are unequal, you can also describe the first case as less than (<) or greater than (>) the second case. As you gain mathematical operations, you gain statistical tools.

Interval/Ratio

The third and final question is “Do the attributes form a numeric scale?” The issue here is, after the attributes are put in rank order, whether the amount of difference between any two adjacent attributes is the same as the amount of difference between any other two adjacent attributes. This property is sometimes referred to as a fixed unit of measurement. And the only way we can precisely say how different one attribute is from its neighboring attributes is if the attributes form a numeric scale.

The attributes of the variable measuring agreement or disagreement with the statement about data analysis certainly have a natural order, but they do not represent a numeric scale. It is not clear precisely how large the difference is between adjacent categories. While moderately disagreeing represents more agreement than strongly disagreeing, how much more agreement does it represent? Nor is it clear that the distance between any neighboring attributes is always the same. For example, is the difference between strongly and moderately disagreeing the same amount of difference as between slightly disagreeing and slightly agreeing? You can’t say because the attributes do not represent a numeric scale. Therefore, this variable has just an ordinal level of measurement.

Now consider a variable that records the number of children a person has. The attributes start at 0 and increase by 1, going as high as necessary for the case with the greatest number of children. This variable’s attributes cover all possibilities without overlapping, can be put in a natural order from low to high, and form a numeric scale. You know precisely how much of a difference there is between any two adjacent attributes: exactly one child.

If the answer to the third question is yes, then you have an interval/ratio variable. Number of children is interval/ratio. With a numeric scale, you gain the mathematical operations of addition (+) and subtraction (–). When two cases are not equal, not only can you state whether the first case is less than or greater than the second case, but you can also subtract the value of the lower case from the value of the higher case to determine how many units apart they are on the numeric scale. Correspondingly, you could add a certain number of units to the value of the lower case to make it equal the value of the higher case. Not surprisingly, the advantages of a numeric scale are substantial for statistical analysis.

SPSS TIP

Caution: When Numbers Don’t Represent Numeric Attributes

You will soon see that many nominal and ordinal variables are represented by numeric codes when entered into SPSS. Just because a variable’s attributes are represented by numeric codes does not make the variable interval/ratio. The attributes themselves, and not just their codes, must correspond to numeric quantities.

Difficult to Determine

With practice, determining the level of measurement of most variables becomes relatively simple. However, a few variables can still be difficult to classify. Even professional researchers disagree about how to treat certain sets of attributes.

A few difficult situations occur with sufficient frequency that guidelines for classifying them are presented here. These may or may not be the same guidelines as your instructor uses. Check with your instructor and follow her or his guidelines.

One commonly occurring difficult situation is when the attributes represent groups of numeric values. Assume you are gathering data about the students in a college data analysis class. Here are two variables each with two different sets of attributes.

- Number of colleges ever attended (1, 2, 3 or 4, 5 to 7, 8 or more)
- Number of colleges ever attended (1 or 2, 3 or 4, 5 or 6, 7 or more)
- Test grade (less than 50, 50 to 59, 60 to 69, 70 to 79, 80 to 89, 90 to 100)
- Test grade (less than 51, 51 to 60, 61 to 70, 71 to 80, 81 to 90, and 91 to 100)

All of these sets of attributes include an **open-ended attribute** or category. For the first two, the open-ended category is at the end; for the last two, it is at the beginning. One of the guidelines used in this text for determining level of measurement is to ignore open-ended categories. Make your decision about level of measurement based on the other categories. Ignoring open-ended categories is usually safe because experienced researchers typically use open-ended categories only where they expect few if any cases.

Even after we have agreed to ignore the open-ended category, however, these four sets of attributes represent still another difficulty. Notice that in the first and third sets of categories, the attributes are not equally wide. For the first set of attributes for number of colleges, the categories include 1 value, 1 value, 2 values, and then 3 values. For the first set of attributes for test grade, all but the last category include 10 values; the last category includes 11 values. In contrast, the categories for the second set of attributes for number of colleges each include exactly 2 values, and the categories for the second set of attributes for test grade each include exactly 10 values. This makes a difference for determining level of measurement. When some or all of the attributes represent groups of numeric values, you need to check the width of the attributes. If they are all equally wide, the variable is classified as interval/ratio, but if the attributes differ in width, the variable is classified as ordinal. (If the categories differ in width, then the distance from the middle of one category to the middle of the next category is not always the same, which means there isn't really a fixed unit of measurement even though the attributes form a numeric scale.) For the four sets of attributes listed above, the first and third represent an ordinal level of measurement, while the second and fourth represent interval/ratio.

This means that even if a variable's attributes represent all the possibilities without overlapping each other, have an inherent order, and represent a numeric scale, the variable might not be interval/ratio. If the attributes are of unequal width, this text classifies the variable as ordinal. A common example of this occurs in regard to measures of income. Surveys will ask respondents to choose from categories such as $0 to $9,999, $10,000 to $29,999, $30,000 to $59,999, $60,000 to $99,999, and $100,000 or more. Because these categories are of unequal width, the attributes represent just an ordinal level of measurement. To specifically avoid level of measurement problems, many researchers prefer not to ask respondents to choose from groups of values. Respondents, on the other hand, often prefer choosing from groups of values—particularly on items that may be sensitive or difficult to recall precisely.

Dichotomies represent still another challenging situation. When it comes to levels of measurement, dichotomies are quite flexible. While normally thought of as nominal variables whose attributes are simply different from one another (e.g., male or female, left-handed or right-handed), they are sometimes seen as also having rank order (e.g., disagree or agree) and are even treated as interval/ratio by some statistical procedures when the dichotomous attributes represent the presence or absence of a particular characteristic. Because dichotomous variables are used in such diverse ways, be sure to ask your instructor how she or he wants you to handle them. This text usually treats dichotomies as nominal-level variables.

CONCEPT CHECK

Without looking back, can you answer the following questions:

- What are some common units of analysis in social science research studies?
- What are variables and attributes?
- What are the three levels of measurement used in this text?
- What are the three questions to ask about a variable's attributes to determine that variable's level of measurement?

If not, go back and review before reading on.

Reliability and Validity

Operational definitions should be both reliable and valid. **Reliability** is often defined as consistency of measurement. An operational definition is consistent or reliable if you get the same scores for cases the second time you measure them as you did the first time you measured them. That assumes, of course, that not enough time has passed between the first and second measurements for the underlying characteristic to have changed. An operational definition would not be very reliable if, when you remeasured subjects, persons whom you first classified as very liberal were now being classified as slightly conservative and if those who were moderately conservative on first measurement were moderately liberal on second measurement. You expect your measurement procedure to put cases in the same attributes or at least nearly the same attribute the second time as it did the first time.

To assess an operational definition's reliability, a researcher may literally retest, re-observe, or re-measure some of the cases and statistically assess how well the scores the cases received the first time match the scores they received the second time. Alternatively, if an operational definition uses not just one question to measure a variable but several questions all aimed at the same underlying concept, reliability can be assessed by examining how well the answers to the different questions match one another. SPSS has several statistics you can use to assess the reliability of an operational definition.

Validity can be briefly defined as measuring what you intend to measure. Validity refers to the goodness of fit between your theoretical definition and your operational definition. Are you measuring what you intend to measure?

Compared to reliability, where you are assessing the match between two sets of scores, validity is more difficult to measure because you are talking about the match between your theoretical definition of the concept and a set of scores produced by your operational definition. Assessing the fit between something abstract (theoretical definition) and something concrete (the scores produced by your operational definition) can only be done indirectly. Sometimes the fit is judged on logical grounds and sometimes on statistical grounds. SPSS can help with the statistical assessment of validity.

Both reliability and validity as described here are characteristics of individual variables. Some (hopefully all) of the variables in a study may have good reliability and good validity, some may have good reliability but poor validity, and still others may have poor reliability and poor validity.

Both the producers and the consumers of research should be concerned about the reliability and validity of the variables used. When variables lack reliability and validity, you can have no confidence in the conclusions of the research. Checking the reliability and validity of operational definitions can be tedious work. It is best done and sometimes can only be done by the researcher who originally gathers the data. When you can't rigorously assess the reliability or validity of a variable's operational definition, at least ask yourself if a researcher's questions, observation techniques, or classifying procedures would likely yield consistent measures if they were repeated (that is reliability) and ask yourself if they are really getting at what the variable is calling itself (that is validity). If either answer is no, the results of the analysis, no matter how good looking, are questionable.

Relationships Between Variables

Sometimes social scientists are interested in just a single variable. For example, how many persons voted for each candidate in an election? More often, however, they look at the relationship between variables. For example, is there a relationship between a person's gender and his or her choice in an election? Much of this text will be about statistically assessing the relationships between variables.

Independent and Dependent Variables

When considering the relationship between two variables, social scientists often think of one variable as influencing the other. For example, there is probably a relationship between the amount of time students spend studying for a course during the first few weeks of the semester and the grade they receive on the first exam. Most people believe that amount of study time influences how well a person does on the exam.

When you are thinking of one variable as influencing the other, the variable that is doing the influencing is called the **independent variable,** and the variable that is being influenced is the **dependent variable.** In the previous example, amount of study time would be the independent variable and exam grade would be the dependent variable.

A variable is independent or dependent only in the context of a relationship. One variable by itself is neither independent nor dependent. Furthermore, the same variable in one relationship may be dependent, but in a different relationship, it may be the independent variable. In the previous example, first exam grade was the dependent variable. However, in the relationship between the variables first exam grade and time spent studying in the weeks following the first exam, first exam grade is the independent variable.

For many of the statistical procedures covered in this book, you will need to distinguish between the independent and dependent variables. It is sometimes difficult to make that determination. Here are some things that can help. First, if one of your variables occurred earlier than the other variable, the variable that occurred first is the independent variable. So, one of the things you can ask yourself is which variable happened first. Which occurs first in a person's life, his or her gender or current marital status? Gender, of course, so gender is the independent variable.

Second, when the time sequence of the variables is not clear, consider if one of your variables represents a broad orientation while the other reflects a more specific opinion or behavior. Usually, broader orientations develop earlier and change less frequently than specific opinions or particular behaviors. For example, which reflects a broader orientation: political party affiliation or preferred candidate in the upcoming election? Since political party affiliation is a broader orientation, it is the more likely independent variable.

For many pairs of variables, you can imagine influence moving in both directions. For example, consider the two variables mother's educational level and daughter's educational level. While daughters could certainly be influenced by their mothers, it is also possible that a mother might be so impressed by a daughter's educational accomplishments that the mother returns to school. The third suggestion for identifying the independent and dependent variables is to consider both directions in which influence might flow and pick the more common or more likely direction. Although some mothers may be influenced by their daughter's educational achievements, the influence of a mother's level of education on a daughter's educational level is the more common path of influence, so mother's education would be the independent variable. Just because it is possible that for some cases, the influence flows in the opposite direction, this should not stop you from identifying the independent and dependent variables. Base your decisions on what is the more likely flow of influence. Only when influence is almost equally likely to flow in either direction should you not identify independent and dependent variables.

Control Variables

In the real world, few dependent variables are influenced by just one independent variable. For example, a grade on a first exam can be influenced not only by amount of time spent studying but also by other factors such as class attendance, year in school, and initial interest in the subject, to name just a few.

Identifying the separate effect of each independent variable on the dependent variable can be tricky because independent variables are sometimes related to one another. For example, you are interested in the effect of study time on test grade. But it is probably true that persons who study more also typically attend class more often. So, when you see that persons who study more tend to get better grades, is it really the effect of study time or is it just the hidden effect of class attendance?

This is where the notion of **control variables** comes in. A researcher will want to control the effect of the other independent variables to better perceive the effect of the one independent variable in which she is primarily interested. Controlling a variable can be done in one of two ways. First, you can control a variable by turning it into a constant. For example, you might limit your study of the relationship between study time and exam grades to only those students who had perfect clas attendance. If you did that and still found that students who studied more got better grades, it could not be because of better class attendance because all the cases in the study had the same level of attendance. The problem with controlling a variable in this way is that you may end up with only a few cases to study, and you would probably want to repeat your study to see if study time affects grades similarly for students with near-perfect attendance, fair attendance, and awful attendance. The second technique to control a variable is to statistically control its effect. Many SPSS procedures can statistically control the effect of other independent variables. Those techniques identify the **net effect** of individual variables on the dependent variable, net effect meaning the effect when the other independent variables have been controlled.

Causality

It takes some pretty strong evidence to support the claim that one variable *causes* another to change. To prove causality, you must demonstrate three things. First, you must show that the two variables are statistically related (sometimes referred to as **covariation**). Second, you must show that changes in the independent variable preceded changes in the dependent variable (sometimes referred to as **temporal sequence**). Third, you must show that the relationship between the two variables is true, that it is still present when you have controlled for all other variables that might be creating the covariation (sometimes known as **nonspuriousness**).

The first requirement of covariation is easy to test. Many of the procedures covered in this book will reveal whether two variables are statistically related. While it is encouraging when the statistical relationship is a strong one, moderate or even weak statistical relationships are sufficient evidence for covariation. All that must be shown is that there is some statistical relationship.

The second requirement of temporal sequence does not depend on any statistical test. Can you be sure, or at least reasonably confident, that what you are claiming is the cause (i.e., the independent variable) happened before what you are claiming is the effect (i.e., the dependent variable)? Sometimes the research design employed by the researcher involves repeated measurements of the same variables over time. Such a design allows a researcher to note when certain changes occur. Sometimes researchers control when subjects are exposed to certain things. That also facilitates determination of temporal sequence. However, in cross-sectional studies, which are conducted at just a single point in time, researchers may have to depend on logic to determine the probable sequence of events. This constitutes the weakest evidence of temporal sequence.

The toughest requirement to meet is nonspuriousness. If something is spurious, it is false. It is not what it appears to be. To prove that the relationship is nonspurious, the researcher must show that the statistical relationship remains when all possible other independent variables have been controlled. The problem is that there are an infinite number of other possible causes for the dependent variable. The statistical techniques mentioned above, which allow you to statistically control the effects of other variables, are helpful but cannot eliminate all other possibilities. If you have controlled for what seem to be the most likely alternative causes and the statistical relationship between the independent and dependent variables remains, you can be reasonably confident that the relationship is nonspurious. Ultimately, though, only the true experimental designs that involve random assignment of subjects to groups and manipulation of the independent variable permit researchers to exclude all other possible explanations and definitively demonstrate nonspuriousness. Regrettably, practical and ethical considerations prohibit the use of true experimental designs to study many important social research questions. That is why researchers turn to quasi-experimental and nonexperimental research designs. While these designs only suggest rather than prove causality, they provide at least some understanding into topics that are ill-suited for experimental research.

All of this talk of causality is to remind you to be humble in describing the results of your data analysis. Claims about one variable being the cause of another should be made only when there is evidence for all the requirements of causality. Be glad when you can speak of moderate and strong relationships in which the temporal sequence is relatively clear and you have controlled for the most likely alternative explanations.

Sampling

Sampling is a complex topic. Just a few of the ways in which it affects data analysis are mentioned here. The topic of sampling is more fully discussed later in the book when statistics are introduced that draw conclusions about populations from sample data.

A researcher typically has some group of entities (perhaps persons, families, organizations, or nations) about which she would ultimately like to make statements. This group is known as a **population.** For some studies, the population of interest might be students currently enrolled at a university; for others, it might be all the households in a particular city or all the states that make up the United States. For example, a population of frequent interest in this book is "1980 young adults," that is, all those persons in 1980 who were residing in the United States and who were in their 20s.

The members of a population are referred to as the **elements** of the population. When a researcher gathers information about each and every element of the population, that researcher has done a **census.** It is important to understand that when you see reference to a census in this book, it does not necessarily mean the Decennial Census of Population and Housing done in the United States every 10 years. Any data set with information on every element in the researcher's population of interest can be referred to as a census.

Obviously, some populations are very large, and conducting a census of such a population would be time, energy, and resource expensive. Fortunately, relatively precise statements about populations can be made with a high level of confidence from certain types of samples.

There are two broad categories of sampling techniques: **probability** and **nonprobability.** In probability sampling, every element in the population must have some chance of being selected into the sample, and it must be possible to actually calculate each element's chance of being selected. Sampling procedures that do not meet these requirements are called nonprobability procedures. Nonprobability sampling techniques include quota sampling, purposive (or judgmental) sampling, snowball sampling, and convenience (or reliance on available subjects) sampling. Nonprobability sampling is often quicker and less expensive than probability sampling, and it is sometimes the only approach possible when going after difficult to identify populations. However, it is not possible to estimate the amount of sampling error in the results of nonprobability samples. That restricts the types of statistical analyses that can be done with the data.

Probability sampling techniques include simple random sampling, systematic random sampling, stratified random sampling, and multistage (or cluster) random sampling. Although often more time-consuming and expensive, probability samples produce data for which it is possible to estimate the extent of sampling error. That opens up additional data analysis possibilities.

As a data analyst, you need to know whether the data set you are using was the result of a census, a probability sample, or a nonprobability sample. Knowing how the cases were selected enables you to correctly choose statistical procedures and allows you to avoid making inappropriate generalizations about the population from which the cases in the data set came.

CONCEPT CHECK

Without looking back, can you answer the following questions:

- What is the difference between reliability and validity?
- What are the differences between independent, dependent, and control variables?
- To show that a causal relationship exists, what three things are required?
- What is the difference between a census and a sample?
- What are some types of nonprobability sampling and some types of probability sampling?

If not, go back and review before reading on.

Important Concepts in the Chapter

attribute
cases
census
constant
continuous variable
control variable
covariation
dependent variable
dichotomy
discrete variable
ecological fallacy
element
independent variable
interval/ratio
level of measurement
net effect
nominal
nonprobability sampling techniques
nonspuriousness
open-ended attribute
operational definition
ordinal
population
probability sampling techniques
reliability
sampling
temporal sequence
theoretical definition
unit of analysis
validity
variable

Practice Problems

1. Indicate how many cases there are in the study and what the unit of analysis is for each of the following:
   a. In a study of individual behavior in public places, a researcher reports that over a period of several weeks, she gathered data on 253 young persons "hanging out" at a mall.
   b. The General Social Survey randomly selects 1,500 adult Americans to interview. Each respondent is asked a number of questions about his values, opinions, and behaviors.
   c. Surveys were sent to 50 rural churches inquiring about their organization, services, and membership.

2. For each of the following, provide a brief theoretical definition and a brief operational definition:
   a. a variable named "self-confidence"
   b. a variable named "social class"

3. You are doing a study of currently enrolled college students. Provide a complete set of attributes for each of the following variables. (Remember that a single attribute can represent a group of different values.)
   a. academic major
   b. percent of classes attended
   c. number of pets owned
   d. satisfaction with college

4. For each of the following variables, indicate if it would be continuous or discrete:
   a. number of Olympic medals an athlete has won
   b. an individual's fastest time in the 100-meter sprint
   c. number of persons living in a household
   d. average household size in a county

5. Regarding level of measurement,
   a. What are the three questions to ask about a variable's attributes to determine that variable's level of measurement?
   b. What sequence of answers to those questions identifies a nominal variable?
   c. What sequence of answers to those questions identifies an ordinal variable?
   d. What sequence of answers to those questions identifies an interval/ratio variable?

6. You are examining a variable's attributes. They represent all the possibilities without overlapping one another, and they have an inherent order. However, they do not form a numeric scale. What level of measurement does the variable have?

7. For each of the following variables, indicate its level of measurement:
   a. a city's average annual rainfall measured in tenths of inches (attributes: 0.0, 0.1, 0.2, . . .)
   b. whether a person voted in the last presidential election (attributes: yes, no)
   c. income last year (attributes: $0, $1 to $9,999, $10,000 to $24,999, $25,000 to $74,999, $75,000 to $149,999, $150,000 or more)
   d. frequency of feeling in the presence of the supernatural (attributes: never, occasionally, frequently, always)
   e. number of days missing from work (attributes: 0, 1, 2, 3–5, 6–10, 11 or more)
   f. age at last birthday (attributes: 0 to 9, 10 to 19, 20 to 29, 30 to 39, 40 to 49, 50 to 59, 60 to 69, 70 to 79, 80 and older)

8. "I never get the same waist measurement twice. The directions always say to relax but they don't say whether to inhale or exhale. Sometimes I inhale and then measure; sometimes I exhale and then measure." Does this statement describe a reliability problem or a validity problem? Explain.

9. If you are concerned that a variable isn't really measuring what the variable's name implies, are you concerned about reliability or validity?

10. For each of the following expressions, identify the independent variable and the dependent variable:
    a. "Absence makes the heart grow fonder." (variables: degree of fondness and distance)
    b. "Too many cooks spoil the broth." (variables: number of cooks and quality of the broth)

11. I am interested in the effect of using a fertilizer on plant growth. Since I know that amount of light will also affect plant growth, I will make sure that the plants receiving the fertilizer get the same amount of light as the plants not getting the fertilizer. In this example,
    a. What is the independent variable?
    b. What is the dependent variable?
    c. What is the control variable?

12. What three things must be shown to prove causality?

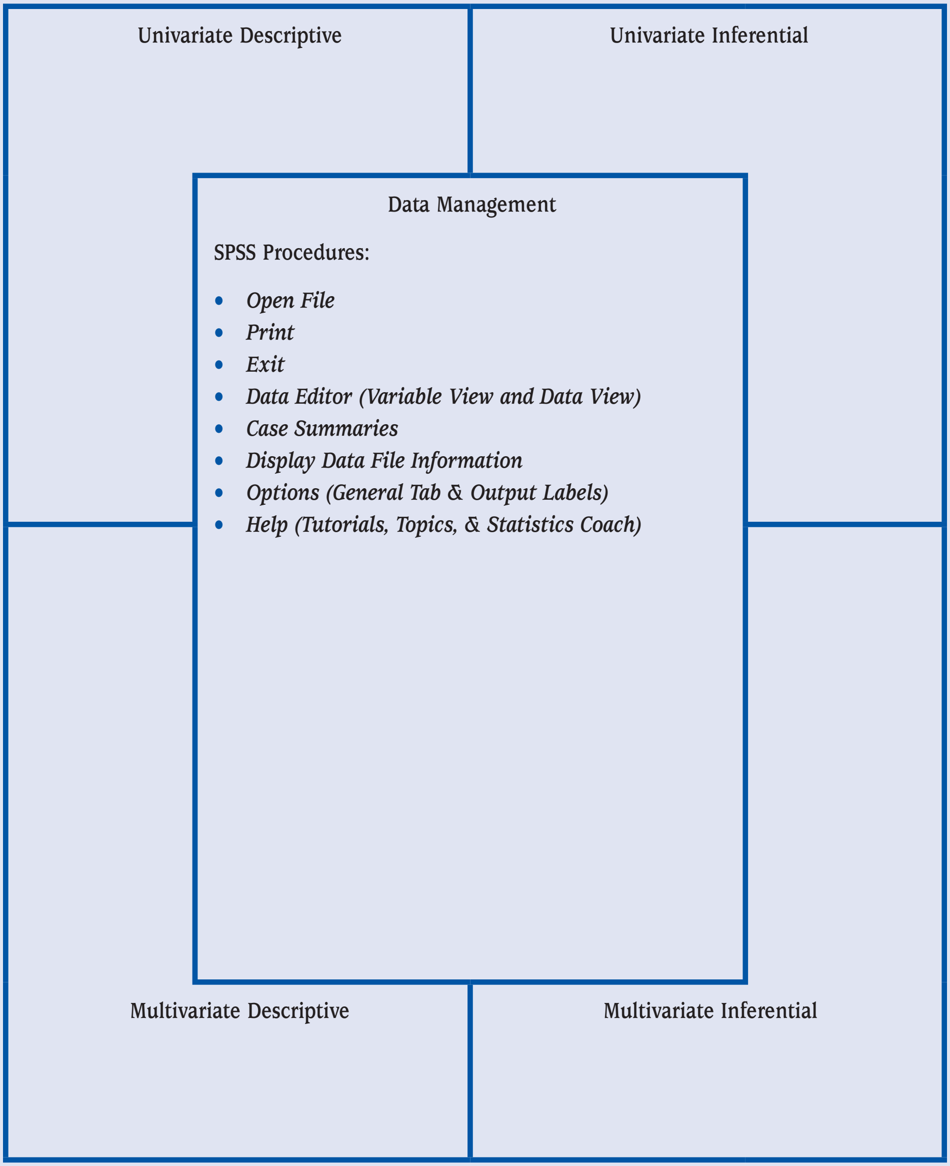
13. A researcher has data for a large number of subjects on the social class of the family in which each person was raised and each person's level of self-confidence at age 20. There is a strong positive correlation between the two variables. What else must the researcher show to prove that being raised in a higher social class causes higher self-confidence?

14. Explain the difference between a census and a sample.

15. Identify for each of the following whether it is a probability sampling technique or a nonprobability sampling technique:
    a. systematic random sampling
    b. quota sampling
    c. snowball sampling
    d. simple random sample
    e. convenience sampling

16. An instructor wants to know how much time the students in his class spent studying for an exam. For each of the following, does the resulting data set represent a census, a probability sample, or a nonprobability sample?
    a. The instructor asks the first five students coming to class the next day.
    b. The instructor randomly selects five students to ask.
    c. The instructor asks each and every student in the class.

Questions and Tools for Answering Them (Chapter 2)



Univariate Descriptive

Univariate Inferential

Data Management

SPSS Procedures:

- *Open File*
- *Print*
- *Exit*
- *Data Editor (Variable View and Data View)*
- *Case Summaries*
- *Display Data File Information*
- *Options (General Tab & Output Labels)*
- *Help (Tutorials, Topics, & Statistics Coach)*

Multivariate Descriptive

Multivariate Inferential