studies conducted at universities, laboratories, and research institutes in the United States must pass the scrutiny of human subjects committees, formally called Institutional Review Boards (IRBs), established under federal guidelines. Such oversight can modify or even disqualify research that is viewed as harmful to the subjects, that does not ensure confidentiality, or in which informed consent of the subjects is not provided. At the international level, a similar protocol applies under the well-known Nuremberg Convention. The latter was specifically created to abolish the kinds of atrocities inflicted upon prisoners by physicians and scientists during the Nazi era.

Another, more subtle disadvantage of the experimental model is the matter of validity. The results that occur under such controlled situations are often affected by the very conditions that make the model useful. Collectively, these conditions are referred to as the *experimental effect*. These range from the "Hawthorne effect," in which the experience of simply being observed affects "after" performance scores, to the "placebo effect," in which the experience of *believing* that one has been subjected to the independent variable can influence outcomes. Another possibility, referred to as the "testing effect," is that participants in an experiment will learn about the expected outcome (such as performing the task of making microchips) from the "before" test, regardless of any influence of the independent variable. These and many more factors have led social researchers to be skeptical about generalizing from experimental outcomes to what people actually do in the real world. A training film might decrease errors in a laboratory, but the workers and managers are not especially interested in that. They want to know what will happen in the factory.

Because of these and related concerns, most social research does not follow the experimental model to the letter. Of course, social scientists appreciate the advantages of the model. But for practical, ethical, and purely scientific reasons, they are more likely to attempt to simulate experimental controls with the use of *statistical* controls and similar techniques. Clearly, the most frequently employed data collection technique is through the use of face-to-face interviews, telephone interviews, and written questionnaires.

Surveys

Each of these *instruments* has distinctive advantages and limitations, but for statistical purposes they are more or less interchangeable. The procedure for their use begins with the researcher constructing the instrument in such a way that the subjects (who may be units of observation, analysis, or both) are prompted to provide data that are relevant to the hypothesis(es) under consideration. These prompts, or items, may be in the form of a question, "Have you ever watched a microchip production training film?" Or, as in the case of a Likert scale, they are statements to which subjects are to indicate their opinion, attitude, or related fact. For example, "Indicate the number of errors per hour your team made before seeing the training film: (A) None, (B) 1 to 5, or (C) more than 5." An appropriate sample is then selected and each subject, or *respondent*, is interviewed or provided with a copy of the questionnaire to complete. The responses are then recorded and summarized for descriptive and/or inductive applications.

The term *survey research* is applied to interview and questionnaire-oriented studies in which the respondents are part of a large, scientifically selected sample. Ordinarily, the instrument is relatively short and the items can be answered quickly. Some items may be *closed-ended*, in which the respondent must select from among a set of predetermined answers, such as those above. Or they can be *open-ended*, which respondents are to answer in their own words: e.g., "How do you feel about watching training films?" But in either case, the subject is not expected to provide a lengthy, personal narrative. Rather, the goal is to gather somewhat superficial responses from a large number of people.

The use of questionnaires and interviews is often equated with survey research, but this is inaccurate. Although surveys are among the most common contexts in which such instruments are employed, long and short instruments—and those that include in-depth probing—are routinely used in experiments and in other kinds of data collection.

Ethnographic Research

The participant-observer or ethnographic approach has been applied in testing all sorts of social scientific hypotheses. With this technique, the researcher chooses to live among and participate in the lives of people, or to become personally involved in situations, under study. This is accomplished not in the laboratory but in natural settings. A very important aspect of this approach is that the researcher keeps a detailed record of his or her observations, usually in the form of field notes or a journal. Much of this record is in the form of firsthand reports: "today I saw . . . yesterday my neighbor . . . ," and so on. But interviews are routinely used to tap into DKAP-type variables as well as to collect information about the life histories or personal experiences of interviewees.

Secondary Data Analysis

In addition to experimentation and interview/questionnaire (including survey and ethnographic) types of data collection, there is a wide range of techniques for which direct contact with human subjects is not pursued. For this reason, they are often referred to as "unobtrusive," to emphasize the fact that they do not disturb or interfere with the lives of those under study. Perhaps the most widely employed among these is *secondary data analysis*. In this case, the researcher does not seek new data but instead bases description and induction on data collected by others. The U.S. Census Bureau and the census operations of other nations are among the most important sources of secondary data, nearly all of which are collected via large surveys. A related and common source is vital statistics on births, marriages, deaths, illness, and similar life-cycle events. In the United States, these are provided by the National Center for Health Statistics (NCHS).

Another type of secondary data consists of documents and firsthand observations maintained in historical and literary archives. These archives often contain items such as personal letters and diaries of well-known, as well as ordinary, people. This kind of data is especially useful to social historians, historical sociologists, and political theorists. Newspapers are also important sources of data in social research, especially (but not necessarily) research on past events. Combining some of the features of officially collected data and documents are such items as sports records, voting data, and economic indicators (e.g., stock market trends). In brief, any information contained in collections, agency records, and the like that already exists in more or less complete form prior to a given research project would be considered secondary data.

Each of the data sources identified here has special characteristics that separate it from the others: experiments emphasize control, questionnaire/interview approaches use the reports of respondents, ethnographic techniques place the researcher in the midst of the research subjects, and so on. However, each ultimately relies on observing (hearing and seeing) and interpreting the behavior, speech, and/or written reports of human actors.

Content Analysis

In contrast, the last category, known widely as *content analysis*, focuses on the products of human activity, treating them as indicators of sociocultural characteristics of the producer. In such research, the unit is generally not a person, group of people, or organization. Rather, it is all or part of a novel, a movie, a painting, a TV program, or similar item that is not meant especially to *inform* but which researchers observe to understand what it *portrays*. Box 3.4 illustrates this approach to data collection.

Regardless of the type of data collection the researcher uses, from the classical experiment to content analysis, statistics has an important role to play. It is perhaps unnecessary to add that statistical tools cannot substitute for sound theoretical reasoning, good research design, or the appropriate choice of observational approaches. However, the ability to describe one's observations and to generalize correctly from them—that is, to apply descriptive and inductive statistics—supports and enhances these other methodological skills. We often associate social statistics with survey research, and with good reason, considering the extent to which developments in each have depended on the other for so many decades. It would be a mistake to equate the two, however, because the field of statistics has a very wide range of applications; and it does not depend on *any* particular way of collecting data.

BOX 3.4

Statistics for Sociologists

Data Collection Using Content Analysis

A study of daytime TV talk shows in which the author of this book was involved illustrates the content analysis procedure. The idea emerged during a casual discussion with a colleague who had spent his youth in the world of carnivals, where both of his parents were employed. We agreed that the old-fashioned carnival "freak show" was more or less dying out, in part because it was considered very offensive. Yet, we speculated, daytime TV now seemed to be fulfilling the same functions—appealing to people's curiosity, satisfying a voyeuristic urge, and so on. This developed into a project in which a sample of daytime talk-show programs would be watched for an extended period and the kinds of guests and problems featured would be recorded. We were ultimately interested in deciding whether an unusually high proportion of time was spent on what could be called "classical freak-show" characters that my colleague recalled from his youth, with such insulting labels as:

- The Fat Man/Lady: extremely obese people
- The Tattooed Man/Lady: people with most of their skin surface covered with tattoos
- The "Hoochie Coochie" Girl: female striptease dancers
- The Thin Man/Lady: people who are extremely underweight
- Bearded Women: women with large amounts of facial hair
- Midgets and Giants: very short or very tall people
- One-Armed Men, etc.: people with missing or malformed body parts
- Sword-Swallowers, etc.: people who can ingest unusual objects

The results of this study indicated that some categories, such as one-armed men, were not represented at all, whereas others—hoochie coochie girls, fat and thin men and ladies—were among the most common types of guests. We also discovered that several other types of "contemporary freak-show" guests and issues were repeatedly featured. These included drug addicts, people who had experienced incest, and prostitutes.

The point of this example is to emphasize that, although many—extremely diverse—people were observed on the TV screens, they were not the units of observation. Instead, our focus was on programs (usually one hour in length) and our main variable was the proportion of each program devoted to "classical freaks." This is typical of the content analysis approach.

Sources of Secondary Data

Researchers typically make the observations on which they base their statistical analyses using one or more of the techniques discussed above. But for the past several decades increasing reliance has been placed on the use of secondary survey data. Because of the difficulty and expense of conducting large-scale social surveys, it is often far more practical to formulate and test hypotheses with data that have already been collected and made available by others. With the help of high-speed downloads and, especially, with the ever-increasing capacity of the Internet, an enormous amount of information is now available and can be accessed with little difficulty by professionals and students. In fact, the data sets used in the SPSS exercises in this book are examples of secondary data.

The most common sources of secondary survey data are the national census operations, now regularly conducted in well over one hundred countries. The sizes of census samples and the number of variables with which they work are so vast that a university or corporation, not to mention a private researcher, cannot manage them. But because the results of census surveys are generally in the public domain, they can easily be used in projects with the most modest budgets. The U.S. Census Bureau has conducted a complete enumeration of the population every 10 years since 1790. In addition, at every census count ever-increasing amounts of data are collected from large samples in every state and county in the nation. This information is now available in hard copy, on CD-ROMs, and online.

The U.S. Census Bureau also provides ongoing estimates of the size and other characteristics of the U.S. population between official decennial counts. These include the monthly current population survey (CPS), which studies such variables as age, household size, occupational characteristics, income, ethnicity, and family structure. The results of the CPS are published in the journal *Current Population Reports* (*CPR*) and online. The Internet address to access the *CPR* and other census data, maps, a complete international database, and news and information about the census operations is: www.census.gov.

Another U.S. government data archive is maintained by the National Center for Health Statistics (NCHS), located in Bethesda, Maryland. The home page for the NCHS web site is at www.cdc.gov/nchwww. Here you will find information on vital statistics, including births, deaths, illness, and related topics. This site contains numerous detailed tables that can be downloaded or printed in whole or in part.

In addition to such official sources, several universities now maintain libraries of data sets collected by social scientists in the course of large-scale government or privately sponsored studies. Among the leading collections in the United States is the one at the University of Chicago's National Opinion Research Center (NORC). NORC is the home of the most widely used database of its type, the General Social Survey (GSS). This study is updated regularly, contains a wide range of variables, and employs a national sample of several thousand respondents. The NORC web site is at www.norc.uchicago.edu.[6]

Similar collections are maintained by the Institute for Research in the Social Sciences (IURC) at the University of North Carolina and the Inter-University Consortium for Political and Social Research (IUCPSR) at the University of Michigan. Each of these organizations makes data sets available to the public, usually for a nominal fee, and each has catalogs available online that describe the sample, the variables, and other features of each set. The relevant web sites are, for the Institute for Research in the Social Sciences, www.iurc.unc.edu, and for the Inter-University Consortium for Political and Social Research, www.icpsr.umich.edu.

Summary

Because data are so central in both major types of applications and, in fact, in all aspects of social statistics, we concluded this chapter with a discussion of the sources of statistical data. On this basis, in the chapters that follow we do not deal directly with the techniques of data collection. There we generally assume that the data we are describing or are employing for inductive purposes already "exist." However, we have in this brief introduction at least provided a framework for understanding that such data do not just appear from nowhere. Rather, a survey, an experiment, a secondary data set, or the like was conducted or accessed to produce the raw material on which we base our analyses.

With this in view, we now move to the part of this book that features the techniques of descriptive statistics. This exploration begins in Chapter 4 with a close and detailed look at the descriptive tool that organizes units, variables, and attributes: the frequency distribution.

KEY TERMS

Bivariate: Hypotheses and statistical techniques that apply to two variables.

Content analysis: A type of unobtrusive research in which books, movies, TV programs, and other cultural artifacts are studied with the aim of understanding those who produced and/or are intended to consume them.

Data collection: Locating and compiling information for descriptive and inductive purposes, synonymous with *scientific observation*.

Dependent variable: The probable or suspected effect in a bivariate hypothesis or application.

Experimental effect: In an experimental design, the change in a dependent variable that is the result of the experiment itself rather than being an authentic effect of the independent variable.

Experimental model: The procedure for gathering data that employs a before/after and experimental-group/control-group design.

Independent variable: The probable or suspected cause in a bivariate hypothesis or application.

Institutional Review Board (IRB): A committee of peer-experts that reviews proposed research involving human subjects. The proposals are ordinarily submitted by colleagues at a university or other research institution.

Instruments: The interview schedules and questionnaires used to gather information directly from the verbal reports of respondents. Instruments are often called "surveys" because of their wide use in survey research.

Interval level: A numerical type of variable whose zero point is an arbitrarily set starting place.

Levels of measurement: A way of classifying variables according to how close or how far their attributes are to being true numbers. The level of measurement determines the range of techniques that can be employed with a given variable. Statisticians generally recognize three or four levels.

Likert scale item: A common ordinal level–type of variable that is often treated as numerical.

Multivariate: Hypotheses and statistical techniques that apply to three or more variables.

Nominal level: The simplest type of variable, whose attributes are names or categories only.

Numerical level: The types of variables whose attributes are actual (cardinal) numbers. Both interval- and ratio-level variables are considered to be numerical.

Ordinal level: The type of variable whose attributes have an inherent rank ordering.

Participant-observer (or ethnographic) approach: A technique of data collection in which the researcher is personally involved (to varying degrees) in the social situation under study.

Ratio level: A numerical type of variable whose zero point literally means complete absence. This is the most complex type.

Secondary data analysis: A type of unobtrusive research in which the researcher works with data that have been collected previously. A familiar application is the use of census data.

Statistical controls: Procedures applied in the collection and analysis of nonexperimental data that simulate the controls that are exercised in an experiment.

Survey research: Interview-and questionnaire-oriented studies in which the respondents are part of a large, scientifically selected sample.

Univariate: Hypotheses and statistical techniques that apply to a single variable.

Unobtrusive research: Any of several research techniques designed so that they do not disturb or interfere with the activities of those under study.

WEB SITES TO BOOKMARK

The following four sites discuss level of measurement.

1. http://trochim.human.cornell.edu/kb/measlevl.htm
   A very complete site supported by Cornell University that includes color graphics.
2. http://courses.csusm.edu/soc201kb/levelofmeasurementrefresher.htm
   This is a refresher discussion with a link to a longer discussion. It is maintained by Professor Kristin Bates at California State University, San Marcos.
3. www.okstate.edu/ag/agedcm4h/academic/aged5980a/5980/chp3/CHAPTER3/sld010.htm
   This is a portion of an online text from Oklahoma State University.
4. www.mors.org/EC2000/presentations/polakoff/sld003.htm
   Here is a research site sponsored by the private organization, Novigen Sciences.

The following three sites focus on independent and dependent variables.

1. www.cs.umd.edu/~mstark/exp101/expvars.html
   This site is maintained by the University of Maryland.
2. www.kmsi.org/curriculum/kmsi/sharedconcepts/data/represent/graph/variables/independent.htm
   This site has interactive graphs and downloadable material.
3. http://www-biol.paisley.ac.uk/scicalc/variables.html
   Here is a brief discussion from England that ends with a surprise.

The following five sources of online data were introduced earlier.

1. www.census.gov
   United States Bureau of the Census.

2. www.cdc.gov/nchs
United States National Center for Health Statistics.
3. www.norc.uchicago.edu
The National Opinion Research Center, University of Chicago.
4. www.iurc.unc.edu
Institute for Research in the Social Sciences, University of North Carolina.
5. www.icpsr.umich.edu
Inter-University Consortium for Political and Social Research, University of Michigan.

SOLUTION-CENTERED APPLICATIONS

1. This application can be done as an individual or a group project. As noted in the latter sections of this chapter, the U.S. Census Bureau is an indispensable source of secondary data to applied social scientists. This exercise introduces you to the vast resources that the Census Bureau now posts online at www.census.gov.

   The allocation of federal funds in the United States is often based on the regional distribution of the population. For example, federal highway funding may be greater for states in some regions than in others. The policies and decisions that affect these kinds of allocations require data on population size and other variables for each of the regions. In this exercise, you are to prepare the kind of data set that is relevant for this type of application. We will use it in several of the solution-centered exercises in this and following chapters.

   The first step is to access the home page of the U.S. Bureau of the Census web site. Explore some of the links to familiarize yourself with the resources that are available. One of the most useful of these links is the "American Fact Finder," whose link is on the left side of the home page. After you have explored this site a little, you are to find values on six variables for the 50 states, recent census data from either 1990 or 2000.

   As one of the variables, select "Region," which is nominal level. The attributes for the main categories of "Region" are Northeast, South, Midwest, and West; and each category is further divided into two or three smaller units such as New England and Mid-Atlantic for the East. Use the larger categories.

   Next, select two variables for which you will create ordinal-level categories.

   - The first of these is "population ages 25 and above with less than a 9th-grade education." The average for the entire United States for this variable is 7.5%. You are to give each of the states a ranking based upon its specific percentage, as follows: *low* = less than 6.0%; *medium* = 6.0% to 8.9%; and *high* = 9.0% and above.
   - The second of these is "percent of population ages 21 to 24 with a disability." The average for the entire United States for this variable is 19.2%. You are to give each of the states a ranking based upon its specific percentage, as follows: *low* = less than 15.0%; *medium* = 15.0% to 23.0%; and *high* = above 23.0%.

   The three other variables will be numerical. One of them will be total population size. The other two are of your choosing and might include number of persons below age 5, percentages in various racial/ethnic categories, etc. With these data, and according to the instructions on creating an SPSS data file in Chapter 2 of the *Study Guide*, create an SPSS data file to be used in later exercises.

2. This exercise focuses on using library resources to identify key statistical concepts. It is a continuation of Solution-Centered Application 2 in Chapter 2. Select three recent issues of sociology journals from the periodical section of your school library. Select one article in each that reports the findings of a research project. For each article, identify the main dependent and independent variables, and also note whether there are other (independent or intervening) variables. Next, determine the level of measurement of each of the variables identified. In a brief write-up, list the name of each article, the author(s), the journal, volume, number, and pages. Then name the variables in each, state their levels of measurement, and provide a brief explanation of your reasons for (a) designating the independent, dependent, and—if appropriate—intervening variables as you did and (b) deciding on the levels of measurement.