IV

DATA PROCESSING, ANALYSIS, AND INTERPRETATION

As with other facets of research, data analysis is very much tied to the researcher's basic methodological approach. In the chapters on field and available-data research, we discussed certain data-analytic techniques at length, but in the case of experiments and surveys we alluded only briefly to this stage of research. The first two chapters in this section take up where our discussions in chapters 7 and 9 left off. Having collected data in a study, the researcher must quantify them, put them in computer-readable form, and analyze them statistically. Chapter 15 charts this process for survey research, outlining the key steps in processing data prior to analysis and describing elementary statistical analyses. Chapter 16 then considers some more advanced statistical techniques for assessing the causal relations among sets of variables.

Even though our attention in these two chapters is focused mostly on the analysis of survey data, the underlying logic and flow of the analysis described are basically the same as in other approaches. There is, for example, always a constant interplay between theory and data. The stage for data analysis is set by the researcher's theoretical model of anticipated relationships because this limits and guides the kinds of analyses that can be carried out. The analysis, in turn, assays and elaborates this model and invariably suggests new models for further analysis. The first rule in all this is that "facts (data) never speak for themselves." Rather, they must be interpreted. That interpretation will be influenced by prior research and ultimately will be communicated in the form of a research report, book, or article that will be read and interpreted by others. Thus, in chapter 17, we examine the reading and writing of research.

15

Data Processing and Elementary Data Analysis

As we consider the later stages of research, it is valuable to recall the broader process of scientific inquiry. Science is a means to understanding that involves a repetitive interplay between theoretical ideas and empirical evidence. Data analysis takes place whenever theory and data are compared. This comparison occurs in field research when an investigator struggles to bring order to, or to make sense of, his or her observations. In surveys and experiments, the researcher typically brings theory and data together when testing a hypothesis once the data have been gathered and processed. In any case, initial data analyses set the stage for a continued interaction between theory and data.

This chapter and the next chapter focus on data processing, analysis, and interpretation. Together the two chapters follow the typical development of a survey study following data collection: from handling the data and putting them in computer-readable form to preliminary analyses involving one and two variables (chapter 15) to more advanced statistical analyses (chapter 16). Both chapters are oriented toward the reader more as consumer than producer of research and in no way can substitute for a course in statistics. In fact, we eschew computational formulas in favor of verbal presentations. What we want to do is give the reader a sense of the *process* of data analysis—a learning process achieved through comparing empirical evidence with theoretical expectations.

Preview of Analysis Steps

In a sense, data analysis begins with a statement of hypotheses, the construction of a theoretical model, or, at the very least, implicitly anticipated relationships among a set of variables; for these models guide the collection of data and thereby determine the alternative relationships or models that may be analyzed. In chapter 4, for example, we discussed Beckett Broh's (2002) study of the impact of extracurricular involvement on academic achievement. In examining this relationship, Broh considered the impact of gender, race–ethnicity, parents' educational attainment, family income, and other antecedent variables. Each of these variables, she reasoned, could either independently affect academic achievement or possibly create a spurious causal relationship between extracurricular involvement and academic achievement. She also identified various intervening variables intended to capture theoretical links between sports participation and educational achievement, which she derived from developmental theory, the leading crowd hypothesis, and social capital

theory. As in all social research, these theoretical expectations guided Broh's selection and measurement of variables and ultimately her analysis of the data.

Chapter 16 focuses on statistical techniques for assessing the causal relations among sets of variables. This is where we see how Broh tested her various theoretical models. Following data collection, however, several steps are necessary to prepare for this final stage of hypothesis testing. Let us preview the basic analysis steps that we cover in this chapter: (1) placing relevant information in computer-readable form, (2) inspecting and transforming the data for the intended statistical analyses, and (3) preliminary hypothesis testing.

In survey research, the first step includes editing and summarizing the responses (coding), data entry, and error checking (cleaning). Some of this **data processing** occurs during data collection in computer-assisted surveys. Researchers performing secondary analysis, as in Broh's analysis of data from the National Educational Longitudinal Survey (NELS), begin their analysis at the second, data-inspection and data-modification, step. The goal of inspection is to get a clear picture of the data in order to determine appropriate statistical analyses and necessary data modifications. Usually, one examines each variable singly (univariate analysis), especially for insufficient variation in responses, missing information, abnormalities, and other weaknesses that may be mitigated prior to the analysis. The reasons for **data modification** are many: For example, a researcher may want to combine the responses to several items in order to create an index or scale, change one or more of the values for a variable, or collapse categories for purposes of analysis. In chapter 5 we described how Beckett Broh created measures of interscholastic sports participation and time spent on homework by combining variables in the NELS data set.

The analysis then turns to empirical testing of the theorized relationships. For simple two-variable (bivariate) hypotheses, the analyst determines if the association between the independent and dependent variables confirms theoretical expectations. Broh, for example, found that participation in interscholastic sports was positively associated with students' math and English grades, as predicted. In a true experiment, assessing the relationship between the independent and dependent variables is the final analysis step because an adequate design effectively controls extraneous variables. But in nonexperimental designs, extraneous variables may pose serious rival explanations that require statistical control. Broh had to explore the possibility, for example, that higher-performing students were more likely than their peers to play interscholastic sports.

Some researchers skip the previous step (preliminary hypothesis testing) in favor of a full-blown multivariate model containing all relevant independent, dependent, antecedent, intervening, and other variables. We prefer to conceptually describe the process as having two interrelated stages. If preliminary hypothesis testing supports theoretical expectations, the analyst formulates multivariate models to rule out, to the extent possible, that the initial results were a spurious consequence of uncontrolled antecedent variables. Conversely, if hypothesized relationships are not supported in preliminary testing, the researcher designs multivariate models to determine if uncontrolled extraneous variables are blocking or distorting the initial results. The preliminary testing step also may reveal unanticipated (serendipitous) findings that suggest alternative multivariate models.

Data Processing

According to James Davis and Tom Smith (1992:60), data preparation (or processing) is "the least glamorous aspect of survey research." Yet, as they also point out, "probably at no other stage . . . is there a greater chance of a really horrible error being made. . . . To avoid such errors, many checks and safeguards must be built into the system." In survey data processing, four essential steps constitute the process of checking the data and making them serviceable for analysis: editing, coding, data entry, and cleaning. Some of these procedures have been taken over by software in computer-assisted telephone interviewing (CATI), personal interviewing (CAPI), or self-interviewing (CASI).

Editing

Editing is a quality-control process applied mostly to paper-and-pencil surveys. Its purpose is to ensure that the information on a questionnaire or interview schedule is ready to be transferred to the computer for analysis (Sonquist and Dunkelberg, 1977). "Ready" means that the data are as complete, error-free, and readable as possible.

Editing is carried out both during and after the process of data collection, and much of it occurs simultaneously with coding (see next section). In interview studies, the editing process begins in the field. Interviewers should check over their completed forms for errors and omissions soon after each interview is conducted. Respondents should be recontacted if necessary, or corrections should be made from memory. Field supervisors may also do some editing at this point, such as determining whether the interviews have been properly conducted from the standpoint of using the correct forms, legibly recording answers, and interviewing the correct respondents.

Most of the editing in large-scale surveys is done in a central office. Here, an editor or coder, who serves much the same function as a copy editor for a writer, goes over each completed form (1) to evaluate interviewers and detect interviewing problems (e.g., inadequate use of probes to obtain answers to open-ended questions); (2) to check for multiple answers to single items, vague answers, response inconsistencies (e.g., reporting 0 hours of television watched in one section and the viewing of a specific TV program in another section), and the like; and (3) to make sure that the interview schedule or questionnaire is complete—that all items, especially those with "missing" responses, have coded values. The editor may bring glaring errors to the attention of the principal investigator or the field supervisor, who provides feedback to interviewers; but mostly he or she simply will make corrections directly on the form.

The NELS involved both on-site and centralized follow-up editing and data retrieval. Students, teachers, and parents completed questionnaires, which were supplemented with data from school officials. Field interviewers checked the student questionnaires at the schools, giving special attention to items designated as "critical," such as a student's sex, race, date of birth, household composition, and parents' employment status and education. If they found missing or undecipherable responses to these critical items, they privately contacted the respondents. For the

other questionnaires and data, similarly designated critical items were checked in the central office and retrieval took place by telephone.

Some editing activities can be programmed into computer-assisted interviewing (CAI) and CASI software. CAI or CASI programs can prompt an interviewer or respondent to answer certain questions, to skip others, to use appropriate response codes, and to review obviously inconsistent responses to related questions. But lacking the intelligence of a human editor, CAI and CASI software cannot determine fully if interviewers and respondents are recording consistent and adequate answers. A Web-based or laptop CASI program, for example, may accept "none of your business" as an adequate answer to an open-ended question.

Coding

Coding for computer analysis consists of assigning numbers or symbols to variable categories. In surveys such as the NELS, the categories are answers to questions; and the common practice, which simplifies data entry and analysis, is to use numerical codes only. For the variable gender or sex in the NELS, a code of 1 was used for males and 2 for females. The particular numbers used are arbitrary; a code of 2 might just as well have been used for males and a code of 1 for females. These numerical codes generally are specified directly on the questionnaire or interview schedule. For example, in the NELS second follow-up survey, students were asked if they had participated in (1) a team sport (baseball, basketball, football, soccer, hockey, etc.), (2) an individual sport (cross-country, gymnastics, golf, tennis, track, wrestling, etc.), or (3) cheerleading, pompom, or drill team. For each question, they circled the highest number that applied among five response alternatives: (1) school does not have, (2) did not participate, (3) participated on a junior varsity team, (4) participated on a varsity team, and (5) participated as a captain/co-captain on any team.

For closed-ended questions, coding is straightforward: There are relatively few categories, and you simply need to assign a different code to each category. For open-ended questions, however, the number of unique responses may number in the hundreds. Coding for this type of question is very much like coding in content analysis (see chapter 12). The researcher tries to develop a coding scheme that does not require a separate code for every respondent or case but that adequately reflects the full range of responses. The idea is to put the data in manageable form while retaining as much information as is practical. (See Box 15.1 for an example of coding open-ended questions.)

If anything, the tendency for novice researchers is to use too few categories, which gloss over potentially meaningful differences. Once the data are coded and data analysis is under way, it is easy to combine code categories for purposes of analysis, but it is impossible to recover lost detail. Consider a questionnaire study that needed a count of respondents' siblings: Instead of asking for that number, respondents were asked to list the ages of any brothers and sisters. Now, one could code only the number of siblings. But consider what happens when all the available information is retained—when the age and gender of each sibling are coded and entered separately into the database. Doing this makes it possible, by computer manipulation, not only to count the number of siblings but also to generate measures

BOX 15.1 Reasons for Giving Up Cigarette Smoking: An Example of Coding Responses to Open-Ended Questions

Developing coding categories for open-ended questions, like many other research activities, involves an interplay between theory and data. Let us follow the procedure used by Bruce Straits (1967) in a study based on personal interviews of ex-smokers ("quitters"), current smokers who tried but were unable to quit smoking ("unables"), and current smokers who have never made a serious attempt to stop smoking ("smokers"). Hypothesizing that specific factors precipitate and support smoking cessation, Straits asked quitters and unables the open-ended question: Why did you want to stop smoking? (Smokers were asked, What reasons might you have for giving up smoking?) Individual responses to the question were too diverse and numerous to manage and analyze without grouping them into a smaller number of categories.

Both theory and data guided Straits's development of coding categories. Theoretical considerations included the hypothesis that current physical ailments (especially those easily connected with smoking) play a more important role in the discontinuation of smoking than health fears for the future (e.g., incurring lung cancer). The coding procedure involved first listing each reason given by the 200 respondents. (Typically studies use a sample of about fifty to one hundered respondents to build codes.) Tally marks were used to note identical reasons. Here is part of the listing:

"I couldn't get over a bad cough" 𝍸 𝍸 ||||

"Job requirements" 𝍸 |

"Not good for health" 𝍸 𝍸 𝍸 |||

"It is too expensive" 𝍸 𝍸 |||

"Live longer" 𝍸 |

"Sore throat" 𝍸 𝍸 |

"To see if I could" 𝍸 𝍸

"Heart attack" |

"Quit for Lent and didn't return" |

"Doctor's advice" 𝍸 𝍸 |||

"Sinus condition" 𝍸 𝍸 |

"Causes lung cancer" 𝍸 𝍸 𝍸 |||

"My wife hates it" |||

Next, coding categories were formed by grouping together reasons that seemed similar from the research perspective. For example, people with bad coughs were placed in the same category as those reporting sore throats. Although the distinction between a cough and a sore throat is important from a medical standpoint, it was not pertinent to Straits's hypothesis.

Empirical considerations also influenced construction of the coding categories. Some reasons had to be lumped together in "other" categories because they occurred too infrequently. An unanticipated category had to be established for quitters who had not made a conscious decision to quit but instead found themselves in situations where they were unable to smoke for a temporary period, such as while recovering from an operation.

The final coding scheme, which was used to code *each* reason given, was as follows.

Current health reasons

1. Didn't feel well (unspecified illness)
2. Nasal congestion (bad cold, sinus, etc.)
3. Cough, sore throat (smoker's cough, dry throat, etc.)
4. Shortness of breath (cutting down on wind, etc.)
5. Poor appetite, loss of weight
6. Other ailments (headaches, nervous condition, nausea, dizziness, etc.)

Advised by doctor

7. Unspecified (ordered by doctor, etc.)
8. Because of specific ailment (heart trouble, palsy, etc.)
9. Doctor said smoking harmful

Future health reasons

10. Smoking is harmful (unspecified)
11. Live longer
12. Fear of cancer (lung cancer, etc.)

Other

13. Sensory dislike (bad taste in mouth, smell, dirty house, etc.)
14. Pressure from close associates (wife, co-worker, etc.)
15. Financial reasons (waste of money, etc.)
16. Test of willpower (to see if could, control life, etc.)
17. Work requirement (smoking not allowed)
18. Quit temporarily (because of operation, illness, Lent, etc., and didn't return to smoking)
19. Other reasons
20. Don't know
21. No answer

The coding scheme represents a balance between too much and too little detail. The classification preserves possibly relevant distinctions such as identifying those who specifically mentioned cancer. The detailed categories are easily collapsed (recoded) when the analysis requires broader groupings. For example, Straits (1967:75) reports that a higher proportion (82 percent) of quitters mentioned current health ailments and/or advice from a doctor than did the unables (60 percent) or smokers (37 percent), who mentioned less immediate health threats or weaker reasons such as "waste of money" or "test of willpower" more frequently.

of number of brothers, sisters, older brothers, younger sisters, respondent's spacing from next sibling, and so on.

Entering the Data

One can think of the data as a matrix or spreadsheet, with observations as rows and variables as columns, which have been entered into a computer file and stored on a disk, tape, CD, or other media.[1] In a data matrix for a survey, the coded responses to each question occupy a designated column in the rows for each respondent. There are several options for data entry. One can enter information from paper-and-pencil surveys into a computer file using data-entry software programmed to detect some kinds of erroneous entries, called computer-assisted data entry (CADE). The NELS used this method to enter data obtained from school administrators and school records. An optical scanner or reader is another option. One device reads pencil-marked "bubbles" adjacent to question response categories on survey instruments or on response sheets, such as the answer forms to the Scholastic Aptitude Test (SAT), and then enters this information into a computer file. The NELS optically scanned the questionnaire data from students, parents, and teachers. Finally, in computer-assisted surveys, interviewers (CATI and CAPI) or respondents (CASI) simply enter the answers directly into a computer laptop or terminal.

Before the development of CAI and CASI direct data entry, optical scanners, and data entry–monitoring software, clerical staff would enter the data by hand using a computer terminal or keypunch machine. This introduced various errors as entry operators misread edits and codes on the survey forms, skipped over or repeated responses to questions, transposed numbers, and so on. The standard procedure for minimizing such data-entry errors is to enter the information on each survey twice (ideally by a different person each time) and then compare the two entries for errors. When punched cards were the means of recording and storing data, this was done on a machine called a "verifier." Nowadays, two persons can independently enter the data into separate computer files, and a software program then compares the two files for noncomparable entries. The same verification strategy can reduce coding errors if each file is independently coded by a different person.

Cleaning

After the data have been entered into a computer file, the researcher should check them over thoroughly for errors. Detecting and resolving errors in coding and in transmitting the data to the computer is referred to as **data cleaning**. This is an essential process that may also identify respondent-related errors. Researchers who have invested a great deal of time and energy in collecting their data do not want their work undermined by avoidable mistakes made at the stage of data processing because, unlike sampling error and certain kinds of measurement error, data-processing errors *are* avoidable. The way to avoid them is to be exceedingly careful about entering the data and to use every possible method of checking for mistakes. Here are a few ways to clean your data.

First, data entry should be verified whenever feasible. When verification is not possible, as in CAI or CASI direct data entry, the error rate may be minimized by careful training and monitoring of the data-entry persons (interviewers or respondents) and extensive field pretesting of the computer-assisted survey procedure.

Beyond training, monitoring, and verification, two cleaning techniques generally are applied. The first of these is sometimes called **wild-code checking** (Sonquist and Dunkelberg, 1977:211). Every variable has a specified set of legitimate codes. The aforementioned NELS questions on sport participation, for example, had possible codes of 1 (for "school does not have") to 5 ("participated as a captain/co-captain"); in addition, a code of 6 was used for "multiple response" and 8 for "missing." Wild codes are any codes that are not legitimate. Wild-code checking consists of examining the values recorded for each item to see whether there are any out-of-range codes, such as 7, 9, and 0 for the sport participation questions. Some CAI, CASI, and data-entry software can be programmed to do this kind of cleaning automatically by refusing to accept wild-code values. Alternatively, one can run a frequency distribution for the variable, check to see if there are any erroneous codes, and then computer-search the data file to find the errors. Of course, one should keep in mind that wild-code checking does not detect typographical or other errors involving legitimate codes.

Third, most large-scale surveys use a process called **consistency checking** (Sonquist and Dunkelberg, 1977:215). The idea here is to see whether responses to certain questions are related in reasonable ways to responses to particular other questions. For example, it would be unreasonable, and therefore an indication of coding error, to find a respondent who is married and age 5 or a medical doctor with 3 years of formal schooling.

One common type of consistency checking involves contingency questions. As you may recall from chapter 10, these are questions designed for a subset of the respondents. The General Social Survey (GSS), for example, first asks people about their employment status last week. If the respondent is working full or part time, he or she is asked the following contingency question: How many hours did you work last week, at all jobs? Consistency checking determines if a question such as this was answered only by those for whom the question was intended. We would not expect an answer to this question if the respondent was unemployed, retired, or in school. To identify such erroneous responses, one must break down the frequency distribution for the contingency question by pertinent subsamples of the population. One might check to see, for example, if any of the subsample of nonworking respondents answered the question on number of hours worked last week.

KEY POINT

Most data-processing errors are avoided with computer-assisted interviewing; to eliminate errors with paper survey forms, data entries should be verified and checked for illegitimate (*wild*) codes and for consistency.

Ideally, a suspected error can be resolved by retracing the process from the survey interview through to the dubious data entry. One may resolve errors by examining the original questionnaires or survey schedules, by listening to taped telephone interviews, or by contacting the original respondents. When this is not

possible, researchers may flag the dubious entries as suspicious or reclassify them as missing data.

One can get an idea of the importance of editing and data cleaning by considering the measures taken to edit and clean the GSS data before the GSS implemented CAI in 2002 (Davis and Smith, 1992). Before the data were entered, interviewers in the field went over their interview schedules after each interview to check for errors; field supervisors reviewed the first two forms submitted by each interviewer to check for completeness, clarity, and following of proper coding procedures and then checked a sample of forms submitted thereafter; and coders or editors at The National Opinion Research Center's (NORC's) central office carried out various checks and prepared the completed forms for data entry. The latter included coding responses to open-ended questions and other responses that were not precoded (e.g., "not applicable" or "no answer") on the form, attempting to obtain missing data on crucial questions, and selecting random cases for validation checks. The data were entered into the computer with the aid of a data-entry program, which automatically alerted the investigator to wild-code errors at the time of entry. Once data from all of the forms had been entered, data from a proportion of the forms were entered a second time to ensure that data entry–errors were minimal. Next, cleaning programs were run, first to check again for wild-code errors and then to do consistency checking for certain questions. When errors were detected, corrections were made and the cleaning process was "repeated until all questions on all cases come up as 'clean.'" Finally, the coders turned this "clean" file over to the researchers, who generated frequency distributions for another round of consistency checking.

Data Matrices and Documentation

Once you have entered all the data into a computer file, you are ready to inspect, modify, and analyze them. Before we discuss these steps, however, let us examine the typical form of data storage, the technical terms, and the documentation required to describe a data set. Imagine that you are preparing the data for secondary analysis. What would others need to know to analyze your data? What documentation accompanies the NELS and other processed archived data?

Two types of information are necessary to adequately describe study data: a matrix of variable values or codes for each observation and supporting documentation. The spreadsheet view in Figure 15.1 illustrates a partial **data matrix** of selected variables (columns) for over 200 countries of the world (rows) gathered from *The World Factbook 2002* and other online sources. The countries are in alphabetical order, and only the first fourteen observations (the As) are shown. The number in each cell of the data matrix represents the value of a particular variable (column) for a particular country (row). The first coded cell value of "27.8" indicates that the 2002 population size (first column) of Afghanistan (first row) is estimated as approximately 27,800,000 persons. One can read down a column to inspect the distribution of values for each variable. The literacy rate in the fourteen countries, for example, ranges from 100 percent (Andorra and Australia) to 36 percent (Afghanistan).

Country	Population Mid-2002 (millions)	Total Fertility Rate	Life Expectancy at Birth (years)			Percent Urban	Literacy Rate (% ages 15 and above)		
			Total	Male	Female		Total	Male	Female
Afghanistan	27.8	5.7	46.6	47.3	45.9	22	36	51	21
Albania	3.5	2.3	72.1	69.3	75.1	43	93	NA	NA
Algeria	32.3	2.6	70.2	68.9	71.7	58	62	74	49
America Samoa	0.1	3.4	75.5	71.1	80.2	53	97	98	97
Andorra	0.1	1.3	83.5	80.6	86.6	92	100	NA	NA
Angola	10.6	6.4	38.9	37.6	40.2	35	42	56	28
Anguilla	0.01	1.8	76.5	73.6	79.5	100	95	95	95
Antigua, Barbuda	0.1	2.3	71.0	68.7	73.4	37	89	90	88
Argentina	37.8	2.4	75.5	72.1	79.0	88	96	96	96
Armenia	3.3	1.5	66.6	62.3	71.1	67	99	99	98
Aruba	0.1	1.8	78.7	75.3	82.2	51	97	NA	NA
Australia	19.6	1.8	80.0	77.2	83.0	91	100	100	100
Austria	8.2	1.4	78.0	74.8	81.3	67	98	NA	NA
Azerbaijan	7.8	2.3	63.1	58.8	67.5	52	97	99	96

FIGURE 15.1. Partial data matrix for countries. *Sources*: *The World Factbook* (http://www.cia.gov/cia/publications/factbook/index.html); Population Division of the Department of Economic and Social Affairs of the United Nations Secretariat, *World Population Prospects: The 2002 Revision and World Urbanization Prospects: The 2001 Revision* (http://esa.un.org/unpp).

The accompanying online documentation explains variable definitions and important data limitations. The total fertility rate (second column), for example, is defined as " . . . a figure for the average number of children that would be born per woman if all women lived to the end of their childbearing years and bore children according to [current birth rates] at each age" (*World Factbook 2002*). Thus, if Angola's 2002 age-specific birth rates do not change over time, women will average 6.4 children during their lifetime. Andorra and Austria, in contrast, currently have very low fertility rates. Perusal of the data for the first fourteen countries suggests research questions for exploring the full data set: Do high-fertility countries tend to have low female literacy? Is there a positive relationship between country literacy levels and life expectancy at birth? Do more urbanized countries have lower birth rates?

Before one tests hypotheses, however, it is important to carefully examine the documentation ("the fine print") for data inconsistencies and other limitations. Even though the *World Factbook 2002* is based on information available on January 1, 2002, some of the data were collected much earlier in time. The literacy rate for Algerian females, for example, is based on a 1995 estimate. Besides data-collection dates, variable definitions may vary by country. The *Factbook* reports that there "are no universal definitions and standards of literacy. Unless otherwise specified, all rates are based on the most common definition—the ability to read and write at a specified age." For the first fourteen countries the specific age was 15, except for Albania (age 9) and Anguilla (age 12). Finally, invariably some of the information

Id	*bys36a*	*bys36b*	*bys36c*	*sex*	*race*	*bypared*	*byfcomp*
124902	3	2	2	1	4	3	1
124915	3	2	2	1	4	3	1
124916	2	3	3	2	4	3	1
124932	2	3	3	2	4	3	1
124939	3	2	2	1	4	3	4
124944	3	3	3	2	4	2	1
124947	2	3	2	2	4	2	1
124966	3	3	3	2	4	3	1
124968	3	3	3	1	4	3	1
124970	1	2	3	1	4	4	1

FIGURE 15.2. Partial data matrix for NELS:88.

is unavailable (missing). For the country data, the missing cells are coded NA (not available). in Figure 15.1.

Figure 15.2 presents a partial data matrix for ten of the 12,578 NELS student respondents in Broh's study. The rows represent respondents, who are identified by unique ID codes (first column). The remaining columns represent variables, whose names may identify time frames (e.g., "by" = base year) and question numbers (e.g., "36a" = question 36a) or abbreviated variable names (e.g., "pared" = parental education). Numerical codes in the matrix cells identify question responses or summary variable categories. For respondent's sex (fifth column), recall that a code of 1 was used for males and 2 for females. Thus the first two listed respondents are males, and the next two are females. For race, a code of 1 was used for "Asian or Pacific Islander," 2 for "Hispanic, regardless of race," 3 for "Black, not of Hispanic origin," 4 for "White, not of Hispanic origin," 5 for "American Indian or Alaskan Native," and 8 for "Missing" (no answer or multiple-race categories were chosen by the respondents). Although Figure 15.2 lists only whites (sixth column), about a third of the students were members of the other race-ethnicity categories.

Once study results are entered into a data matrix or equivalent computer-readable format, the researcher should prepare the **codebook** documentation. A codebook is like a dictionary in that it defines the meaning of the numerical codes for each named variable, such as the NELS codes for sex and race. Codebooks also may contain question wording, interviewer directions, and coding and editing decision rules, such as how to handle two answers circled to a single-response question. Originally, survey codebooks were single physical documents, but now this information may be scattered across various sources and media. In addition to an electronic codebook accompanying NELS data sets on CDs, numerous technical and methodological reports are available online from the National Center for Education Statistics (http://nces.ed.gov/surveys/NELS88/). The NELS electronic codebook reports, for example, that sex was coded from the "Your Background" section of the base year student questionnaire, or, when unavailable from this source, from school rosters. If sex was missing from both the student questionnaire and the school ros-

ter, it was imputed from the student's name if this could be done unambiguously. Extremely complex coding rules are given for parental education (seventh column of Figure 15.2) and family composition (eighth column), which are composite variables based on relevant responses from student and parent questionnaires.

When variables are coded directly from a questionnaire, the codebook entries look very much like the survey instrument with the addition of response and nonresponse codes. The entry for Broh's parent–child social capital variables (columns 2–4 of Figure 15.2) includes:

> 36. Since the beginning of the school year, how often have you discussed the following with either or both of your parents or guardians?
>
> {bys36a} Selecting courses or programs at school
> {bys36b} School activities or events of particular interest to you
> {bys36c} Things you studied in class

Code	Label
1	NOT AT ALL
2	ONCE OR TWICE
3	THREE OR MORE TIMES
6	{MULTIPLE RESPONSE}
8	{MISSING}

A code of 6 indicates students who marked more than one response on the optical scan questionnaire. If they failed to respond at all, a code of 8 was used to indicate missing information. In many surveys, missing data for a variable are divided into three categories of those who "don't know" (DK), who provide "no answer" (NA), or for whom the question is not applicable (NAP).

An excellent online site for viewing codebook documentation for the GSS is maintained by the NORC (http://www.norc.org/GSS+Website/Documentation/). The GSS, which we described in chapter 9, consists of interviews administered to national samples using a standard interview schedule (Davis and Smith, 1992). Among the items covered are (1) requests for standard background information on gender, race, occupation, religion, and so on; (2) questions about the quality of the respondent's life in various areas (e.g., health, family, social relations); (3) requests for evaluations of government programs; and (4) questions covering opinions and attitudes about a wide range of issues, such as drug usage, abortion, religious freedom, and racial equality. One can search the online codebook for all GSS variables, either by alphabetically arranged variable name ("abany" to "zombies") or by subject ("abortion" to "world war"). Information for each GSS variable includes question wording, response codes, response trends over time, links to related appendices and methodological reports, and an annotated bibliography of previous usage.

The Functions of Statistics in Social Research

Starting with a cleaned data set, the next analysis step would be to inspect the data to decide on subsequent statistical analyses. But how do you decide what kind of

analysis to do? Naturally, that decision depends on what you want to know. Broh wanted to know if playing interscholastic sports (the independent variable) promotes academic achievement (the dependent variable). Of course, to establish a causal relation such as this, she would need to show that the two variables are associated (that changes in one variable accompany changes in the other), that the direction of influence is from the independent variable to the dependent variable, and that the association between the variables is nonspurious. To begin her analysis, Broh could have examined the NELS data respondent by respondent to see if playing interscholastic sports was associated with academic achievement. But with over 12,000 respondents, this task would be incredibly tedious and probably not very reliable. Therefore, what she needed was a fast and efficient means of summarizing the association between these variables for the entire sample of respondents. This is precisely where statistics (and computers) enters in. A *statistic* is a summary statement about a set of data; statistics as a discipline provides techniques for organizing and analyzing data.

Traditionally, statistics has been divided according to two functions—descriptive and inferential. **Descriptive statistics** is concerned with organizing and summarizing the data at hand to make them more intelligible. The high and low scores and average score on an exam are descriptive statistics that readily summarize a class's performance. Broh needed a descriptive statistic that would summarize the degree of association between playing interscholastic sports and academic achievement in the NELS sample. **Inferential statistics** deals with the kinds of inferences that can be made when generalizing from data, as from sample data to the entire population. Broh needed this form of analysis as a means of determining what the NELS sample of observations indicated about the effects of playing interscholastic sports on academic achievement in the target population of American secondary students.

Based on probability theory, inferential statistics may be used for two distinct purposes: to estimate population characteristics from sample data (discussed in chapter 6) and to test hypotheses.[2] Traditionally, hypothesis testing is employed to rule out the rival explanation that observed data patterns, relationships, or differences are due to chance processes arising from random assignment (in the case of experiments), sampling error (in the case of probability sampling), random measurement error, or other sources.[3] In experiments, such testing determines if the differences between experimental conditions are "statistically significant"—that is, not attributable to random assignment. The appropriate role of significance tests in the analysis of nonexperimental data is a longstanding controversy (see Morrison and Henkel, 1970). Following David Gold's argument (1969) that results that easily could have occurred by a chance process should not be taken seriously, we view tests of significance as an effective means of screening out trivialities and chance mishaps.

Inspecting and Modifying the Data

At this stage, the researcher is ready to give the computer instructions for inspecting and transforming the data.[4] Some of these operations may occur during earlier data-cleaning and coding processes. The goal of inspection is to get a clear picture

of the data by examining one variable at a time. The data "pictures" generated by **univariate analysis** come in various forms—tables, graphs, charts, and statistical measures. The nature of the techniques depends on whether one is analyzing variables measured at the nominal/ordinal level or variables measured at the interval/ratio level (discussed in chapter 5). Following data inspection, the researcher may want to change one or more variable codes, rearrange the numerical order of variable codes, collapse variable categories, impute estimated values for missing data, add together the codes for several variables to create an index or scale, and otherwise modify the data for analysis.

Nominal- and Ordinal-Scale Variables

Broh might have done univariate analyses simply to get a sense of the nature of the variation in the variables to be analyzed. It is generally a good idea, for example, to see if there is sufficient variation in responses to warrant including the variable in the analysis. As a rule, the less variation, the more difficult it is to detect how differences in one variable are related to differences in another variable. To take the extreme case, if all students scored the same on self-esteem, then it would be impossible to determine how *differences* in this variable were related to differences in academic achievement. Fortunately, there was sufficient variation in Broh's main independent variable as nearly a third of the students reported participating in interscholastic sports during both the 10th and 12th grades.

A core GSS question asks, "Would you say that most of the time people try to be helpful, or that they are mostly just looking out for themselves?" Suppose that we wanted to inspect the responses to this question in the 2000 GSS. One means is to organize responses into a table called a **frequency distribution**. A frequency distribution is created by first listing all the response categories and then adding up the number of cases that occur in each category. If we instruct the computer to do this, our output might look like that in Table 15.1, which gives the 2000 distribution of responses to the helpfulness question. This certainly presents a clearer picture than does a case-by-case listing of responses. However, this sort of table generally would

TABLE 15.1. Frequency Distribution of Belief in the Helpfulness of People,* 2000 General Social Survey

Code	*Label*	*Frequency*
1	Try to be helpful	866
2	Just look out for themselves	851
3	Depends (volunteered)	166
7	Don't know	10
8	No answer	3
9	Not applicable	921
Total		2817

*Question: "Would you say that most of the time people try to be helpful, or that they are mostly just looking out for themselves?."

TABLE 15.2. Percentage Distribution of Helpfulness of People for the Distribution in Table 15.1

Response	*%*
Try to be helpful	45.7
Just look out for themselves	45.0
Depends (volunteered)	8.8
Don't know	0.5
Total	100.0
(Number of responses)	(1893)
(Missing data)	(924)

serve as a preliminary organization of the data because more readable formats can be derived from it. In particular, researchers often compute the percentage of respondents in each category. To see how this might create a still clearer picture, examine the raw figures in Table 15.1.

Notice that the number of "try to be helpful" responses in the sample is 866. This number by itself is meaningless unless we provide a standard or reference point with which to interpret it. More than likely as you peruse the table you will see the figures in relation to one another, invoking implicit points of comparison. You may note, for example, that there are slightly more "try to be helpful" than "just look out for themselves" answers or that 166 respondents volunteered a qualified response ("it depends"). **Percentage distributions** provide an explicit comparative framework for interpreting distributions. They tell you the size of a category relative to the size of the sample. To create a percentage distribution, you divide the number of cases in each category by the total number of cases and multiply by 100. This is what we have done in Table 15.2. Now you can see more clearly the relative difference in responses. Now we see, for example, that fewer than 9 percent of the respondents volunteered a qualified response.

It should be noted that the percentages in Table 15.2 are based on the total number of responses, excluding "missing data"—those in the "no answer" and "not applicable" categories. About one-third (921) of the respondents were in a random subsample that were not asked the helpfulness question, and three others (the "no answers") either were not asked the question (interviewer error) or did not respond. Since these are not meaningful variable categories (i.e., they say nothing about belief in the helpfulness of people), it would be misleading to include them in the percentage distribution. The total number of missing responses is important information. If this information is not placed in the main body of a table, then it at least should be reported in a footnote to the relevant table or in the text of the research report. Also notice that the base upon which percentages are computed (1893) is given in parentheses below the percentage total of 100 percent. It is customary to indicate in tables the total number of observations from which the statistics are computed. This information may be found elsewhere—at the end of the table title or in a headnote or footnote to the table; often it is signified with the letter *N*.

Univariate analysis is seldom conducted as an end in itself, especially in explanatory research. One important function mentioned earlier is to determine how to collapse or recode categories for further analysis. For the helpfulness variable, one might limit the analysis to those who directly answered the question (codes 1 and 2 in Table 15.1), but this would result in the loss of 176 "depends" and "don't know" responses that then would be recoded as "missing" for analysis purposes. Instead, one might retain the "depends" and "don't know" responses by recoding them as middle responses in a three-category ordinal scale:

1 Try to be helpful
2 Depends, Don't know
3 Just look out for themselves

Alternatively, one could collapse the responses into a dichotomy by comparing either of the extreme categories with all others. Respondents definitely expressing low faith in people, for example, could be contrasted with the less pessimistic or uncertain responses:

1 Just look out for themselves
2 Depends, Don't know, Try to be helpful

Either of the above recodes achieves the practical objective of minimizing the "missing" observations. The choice of one of these, or an alternative recoding, should be guided primarily by theoretical considerations.

A univariate inspection also can inform decisions about how to collapse the categories of a variable for further analysis. Collapsing decisions may be based on theoretical criteria and/or may hinge on the empirical variation in responses. Thus, years of education might be collapsed into "theoretically" meaningful categories (grade 8 or lower, some high school, high school graduate, some college, college graduate) on the basis of the years of schooling deemed appropriate over time in the United States for leaving school and qualifying for certain occupations. Alternatively, one might collapse categories according to how many respondents fall into each category. If the sample contains only a handful of respondents with less than a college education, these respondents may be placed in one category for purposes of analysis.

In the absence of theoretical criteria for collapsing, the best strategy is to try to obtain an approximately equal proportion of cases in each category. If the distribution is to be dichotomized, this would mean a 50:50 split. Achieving such a split for a nominal variable is a matter of combining conceptually similar categories. For example, if we wanted to compare students with different majors on some relevant variable and we found that 50 percent majored in mathematics or natural science, 30 percent in sociology, and 20 percent in psychology, then we might combine the latter two categories. At the ordinal, interval, and ratio

KEY POINT

A *frequency distribution* can indicate whether there is enough variation to include a variable in the analysis and how to recode or collapse categories for further analysis.

levels, a variable is typically collapsed into about five to eight categories in order to preserve essential information from the original distribution.

Information is invariably lost when the original categories of a variable are collapsed. Ideally, the information lost should not be pertinent to the intended research. One might dichotomize the age of Cuban respondents into those older than or younger than age 60, for example, if the only research purpose was to compare older Cubans with those too young to have experienced pre-Castro Cuba. Similarly, respondents with 9, 10, or 11 years of education may be viewed as essentially equivalent ("some high school") in studying occupational attainment. For other purposes, such as health conditions and practices, dichotomizing age or collapsing education into a few categories would result in a serious loss of vital information.

Interval- and Ratio-Scale Variables

Creating frequency or percentage distributions is about as far as the univariate analysis of nominal- and ordinal-scale variables usually goes. On the other hand, data on interval and ratio variables may be summarized not only in tables (or graphs) but also in terms of various statistics. Consider a core GSS question measuring television viewing, which asks respondents "On the average day, about how many hours do you personally watch television?" Since respondents' answers were recorded in number of hours (the range was 0–24), this variable may be considered a ratio-scale measure. We could get a picture of the number of hours watched, as we did with the helpfulness variable, by generating a distribution of the responses. Table 15.3 presents a computer-like output for this variable for the 2006 GSS. Though not a problem here, the relatively large number of values for most interval and ratio variables makes it necessary to collapse categories to get a compact, readable table. Had we generated the distribution for the variable of age, for example, we would have had so many values—over seventy—that they might not have fit on a single page of computer output. In this case, we might lump together respondents according to the first age digit—those under 20, between 20 and 29, 30 and 39, and so on.

Notice that Table 15.3 presents two kinds of distributions: frequency and percentage. Can you tell from the table what percentage of respondents report that they do not watch any television at all on the average day? We also could get a picture of a distribution by looking at its various statistical properties. Three properties may be examined. The first consists of measures of central tendency—the mean, median, and mode. These indicate various "averages" or points of concentration in a set of values. The **mean** is the arithmetical average, calculated by adding up all of the responses and dividing by the total number of respondents. It is the "balancing" point in a distribution because the sum of the differences of all values from the mean is exactly equal to zero. The **median** is the midpoint in a distribution—the value of the middle response; half of the responses are above it and half are below. You find the median by ordering the values from low to high and then counting up until you find the middle value. The **mode** is the value or category with the highest frequency. The modal value in Table 15.3 is 2 hours. With the aid of a computer program we calculated a mean of 2.78 hours and a median of 2.27 hours.

TABLE 15.3. Number of Hours of Television Watched on Average Day, 2006 General Social Survey

Code	*Label*	*Frequency*	*Percentage*
00	0 HOURS	87	4.4
01	1 HOURS	459	23.1
02	2 HOURS	578	29.1
03	3 HOURS	343	17.3
04	4 HOURS	217	10.9
05	5 HOURS	119	6.0
06	6 HOURS	87	4.4
07	7 HOURS	20	1.0
08	8 HOURS	32	1.6
09	9 HOURS	3	.1
10	10 HOURS	21	1.1
12	12 HOURS	11	.5
14	14 HOURS	5	.2
15	15 HOURS	1	.1
18	18 HOURS	1	.1
24	24 HOURS	1	.1
98	DON'T KNOW	4	MISSING
99	NO ANSWER	2	MISSING
TOTAL		1991	100.0
(VALID CASES) (1985)		(MISSING CASES)	(6)

A second property that we can summarize statistically is the degree of variability or dispersion among a set of values. The simplest dispersion measure is the **range**. Statistically, this is the difference between the lowest and highest values, but it is usually reported by identifying these end points, such as "the number of hours of television watched ranged from 0 to 24 hours." Of several other measures of dispersion, the most commonly reported is the **standard deviation**. As we saw in chapter 6, this is a measure of the "average" spread of observations around the mean. One of its important uses in statistics involves the calculation of "standard scores." A standard score is calculated by dividing the standard deviation into the difference between a given value and the mean of the distribution. This converts the value from a "nonstandard" measurement (e.g., hours, years) specific to the variable and sample to a "standardized" measure expressed in terms of standard deviations from the mean. Similar in some ways to "percentaging" a table, standardizing provides a reference point for comparing individual responses in the same or different distributions.

With respect to the variable of television viewing, the standard deviation is perhaps best interpreted as an index of heterogeneity, which could be used to compare the degree of variability in hours watched among different subsamples or in sam-

ples from different populations. The standard deviation of the foregoing distribution of all GSS respondents was 2.15. Among respondents with less than a high school education, the standard deviation in hours watched was 2.73, revealing more variability in this group than among those respondents with at least some college, for whom the standard deviation was 1.88. As a further example, the ages of GSS respondents in 2006 ranged from 18 to 89, with a standard deviation of 16.55 years. By comparison, the standard deviation for age in a sample of college undergraduates would be around 1.5 years.

A third statistical property of univariate distributions is their shape. This property is most readily apparent from a graphic presentation called a **frequency** or **percentage polygon**. Figure 15.3 presents the percentage polygon for the data in Table 15.3. The figure reveals that the distribution has a single high point (or mode), with the data lopsided or "skewed" mostly to the right (or positive side) of this point. This shape also is typical of income distributions. Many variables in social research have "bell-shaped" distributions, so called because they form the general shape of a bell. In a bell-shaped distribution, the three measures of central tendency are identical, whereas in a positively skewed distribution like Figure 15.3 the mode has the lowest value (2), followed by the median and then the mean. One particular type of bell-shaped distribution, called the **normal distribution**, describes the shape of many variables and statistics, such as the sampling distribution of the mean. The normal distribution is a very important concept in inferential statistics.

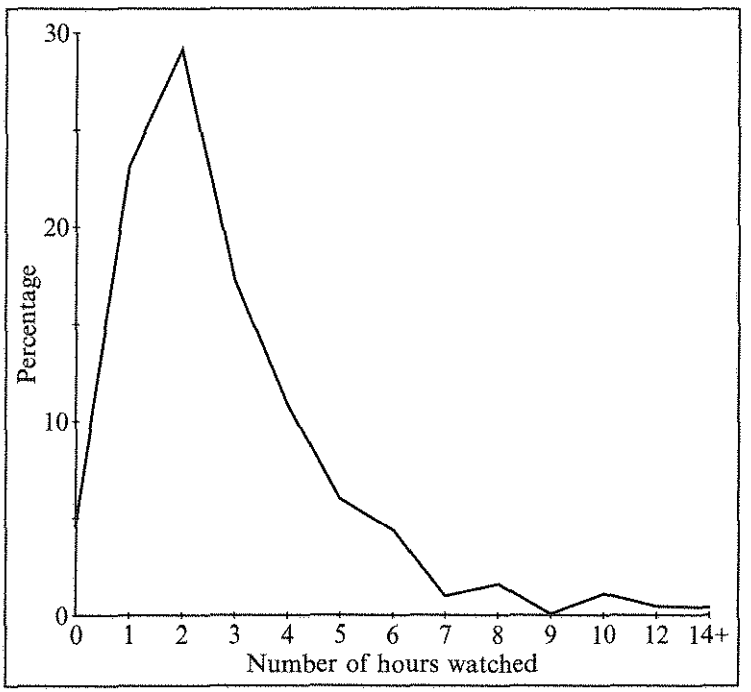

FIGURE 15.3. Percentage polygon for the distribution in Table 15.3.

Collectively, these three statistical properties—central tendency, dispersion, and shape—provide such a good picture of quantitative data that they often obviate the need for tabular presentations. Many investigators, in fact, describe their data simply in terms of a mean or median, an index of dispersion, and occasionally the overall form (for which there are also statistical indices). Other researchers prefer to be as "close" as possible to the data and use various graphical procedures to inspect the data for unexpected patterns and other anomalies (see Box 15.2). If a researcher, for example, computes only the mean and standard deviation of reported television viewing and fails to inspect the entire distribution (Table 15.3), he or she would miss the exaggerated report of 24-hour and possibly 18-hour daily television watching. It is important to spot extreme values or **outliers** as they can adversely affect some statistical procedures and may have to be excluded from the data analysis.

KEY POINT

The distributions of interval/ratio variables can be described in terms of their central tendency, variation, and shape.

One important function of data inspection is to determine the prevalence of missing values. The simplest way to handle cases with missing values, which we did in percentaging Table 15.2, is to remove them from the statistical calculations. This method, called **listwise deletion**, often is used when there are relatively few missing cases. Excluding cases with missing data on *any* of the variables in a planned multivariate analysis, however, can lead to a much smaller, biased sample that is unrepresentative of the target population. Paul Allison (2002) relates a hypothetical illustration of a sample of 1000 that shrinks to only 360 respondents after excluding the 5 percent of missing values on each of the twenty variables.[5] Consequently, researchers typically use various ad hoc or formal procedures for replacing or otherwise handling missing data.

When missing data result from the respondents' uncertainty ("don't know") or refusal to answer the question ("no answer"), a recoding operation, as previously described for the GSS helpfulness question, may eliminate the missing values. Alternatively, researchers with in-depth knowledge of the data may use their expertise to replace missing values with "educated guesses." Suppose that one wanted to study the male literacy rates for the countries illustrated in Figure 15.1. One might replace the missing values for males with the country's total literacy rate or with a corresponding male rate for a similar, neighboring country. In like manner, other responses within a survey interview, or from a similar respondent, may be used on an ad hoc basis to replace missing data.

Besides these ad hoc procedures, many formal statistical solutions, called **imputation**, have been devised to replace missing values with a typical value calculated from the available ("nonmissing") data. Broh simply replaced her NELS missing values with variable sample means. According to Allison (2002), however, this method often produces biased statistics and generally should be avoided. Consider, for example, replacing missing data on respondents' ages with the sample mean for the available data. Although the replacements will not change the calculated mean age, they will bias downward the estimated standard deviation as all imputed values are zero distance from the mean. The most sophisticated imputation procedures for handling

BOX 15.2 Television Viewing and Labor Force Status: An Example of Graphing Data

Developments in graphic procedures for data analysis (Cleveland, 1985:1), when implemented in computer software, offer an attractive alternative to strictly numerical analysis of data. One useful tool is the *box-and-whisker* or *box plot* (Tukey, 1977:39–41). Figure A presents box plots that summarize television viewing reported by the 1985 GSS female respondents grouped according to their current labor force status. (The personal computer program Stata produced these examples.)

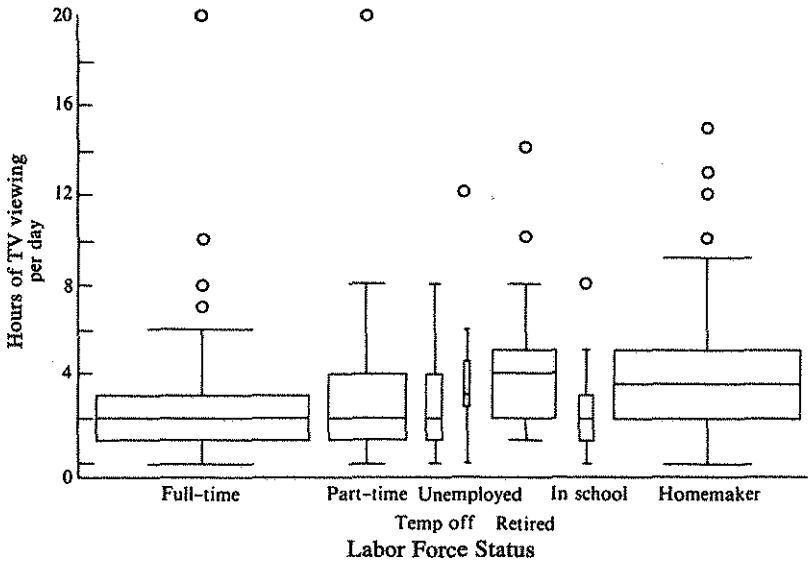

FIGURE A. Average daily television viewing by labor force status, female, 1985 General Social Survey.

The box plots aptly display key aspects of each frequency distribution. The top and bottom of each box indicates, respectively, the 75th and 25th percentiles of television viewing, and the crossbar within each box shows the median (50th percentile). The height of a box is one measure of dispersion, showing the spread of television viewing among the middle 50 percent of the respondents (e.g., 1–3 hours for full-time female workers). An off-center position of the crossbar (median) within the box reflects skewness in the central 50 percent of the distribution (e.g., positive skewness for part-time workers). The width of each box is proportional to the number of respondents; this serves to flag patterns in the data that must be ignored or cautiously interpreted because of small sample sizes (e.g., the unemployed).

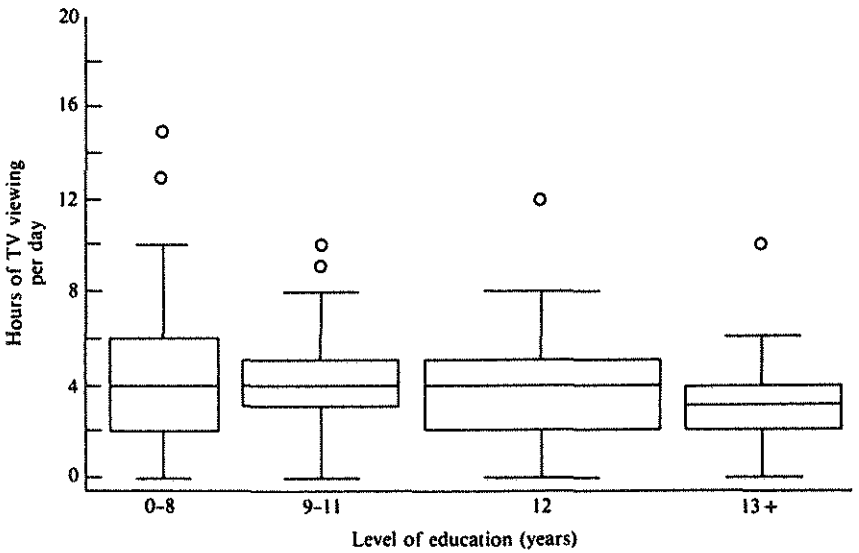

FIGURE B. Average daily television viewing by education, female homemakers, 1985 General Social Survey.

The vertical lines (or "whiskers") extending above and below each box include all but the most extreme data values, which are plotted separately as circles.* The whiskers serve to fence out stray values, which are possibly *outliers* requiring special attention. Sometimes outliers result from measurement error, which seems likely here for the full-time worker who watches television 20 hours a day. Conversely, the reports of watching television 8 hours daily by a few full-time workers are not overly suspicious. Extreme outliers are usually excluded from the data analysis (especially if measurement error is suspected) or analyzed separately from the main data since some statistical procedures may be adversely affected by a few unusual values.

Now you should be able to intepret the box plots in the figure. Which groups report the lowest median television viewing? Which has the lowest variation (dispersion) in viewing habits? How do the retired differ from those keeping house?

Box plots are one of a set of tools designed for exploratory data analysis, which has applications ranging from preparing data for testing prior hypotheses to dredging data for theoretically unexpected patterns. Because of their exploratory purpose, these techniques are relatively resistant to outliers and other data anomalies. And with an appropriate computer program it is quite easy to explore a data domain. When you studied the box plots in the figure, for example, you may have wondered about the large variation in television viewing among women working at home. Perhaps the viewing differences among the homemakers reflect, in part, educational differences. This hunch is easily checked by requesting the computer program to draw box plots for various educational groups (8 years or fewer of completed schooling, 9–11, 12, and 13 years or more), which is shown in Figure B. What do these box plots show?

*Each whisker extends 1.5 times the height of the box or to the most extreme value if that distance is shorter.

missing data predict missing values from the known values of other variables. These computer-intensive methods, which are beyond the scope of this textbook, are difficult to learn and implement and are underutilized in survey research; nonetheless, they offer the best solutions for the missing data problem.

Another important function of data modification is to reduce data complexity by combining variables into indexes, scales, or other composite measures. We discussed this process in relation to both measurement and the use of multiple methods in earlier chapters. The simplest composite measure would consist of combining two "yes/no" questions. This would produce four distinct response categories: (1) yes, yes; (2) yes, no; (3) no, yes; and (4) no, no. An index based on the number of "yes" responses reduces the category total from four to three (0, 1, or 2 yes's). Similarly, the sixteen possible response patterns for four "yes/no" questions reduces to five by indexing the affirmative responses (0, 1, 2, 3, or 4 yes's). Paul Lazarsfeld (1972) describes this data-reduction approach as "arbitrary numerical" as formerly distinct categories (e.g., "yes, no" and "no, yes") are combined.

Beckett Broh used this approach to create measures of several concepts. For example, she measured social capital between students and parents by combining responses to three questions (see p. 508), which asked how frequently students talked with their parents about course selection, school activities, and studies. Possible responses included "not at all," "once or twice" and "three or more times," which she coded as 0, 1, and 2, respectively. Thus, when these questions were summed, the composite measure of "student-parent talk" had a possible range of 0 to 6.

Finally, the assignment of numerical values to the categories of nominal-scale variables may be used to transform them into ordinal or interval scales for some purposes. In the GSS, for example, the occupations of respondents and their relatives (mother, father, spouse) are coded into 503 U.S. Census occupational categories. Numerical values representing occupational characteristics, such as prestige ratings or gender composition, may be used to convert the 503 nominal categories into simpler, continuous variables for studying, say, occupational mobility or gender segregation. Each coded occupation in the GSS is assigned a prestige score, which ranges from 86 (physicians) to 17 (miscellaneous food preparation occupations).[6] Occupational complexity is substantially reduced as formerly distinct occupations receive the same prestige ratings, such as prestige scores of 73 (architects, chemical engineers, and chemists), 49 (funeral directors, real estate sales occupations), and 22 (messengers, janitors and cleaners, hand packers and packagers). Similarly, the percentage female in each Census occupational category may be used to transform the 503 categories into a continuum ranging from about 2 percent (e.g., mechanical engineers) to 96 percent (kindergarten teachers) for studying gender-based occupational segregation (Straits, 1998).

Preliminary Hypothesis Testing

Following data inspection and necessary modifications, researchers turn to empirical testing of bivariate relationships. The object of **bivariate analysis** is to assess the relationship between two variables, such as that between extracurricular involve-

ment and academic performance. In general, this amounts to determining, first, whether the relationship is likely to exist (or whether it might be a product of random error) and, second, how much effect or influence one variable has on the other. As with univariate analysis, the way in which this is done depends on the level of measurement.

Nominal- and Ordinal-Scale Variables

When the variables analyzed have only a few categories, as in most nominal- and ordinal-scale measurement, bivariate data are presented in tables. The tables constructed are known as cross-tabulations, cross-classifications, or contingency tables. A cross-tabulation requires a table with rows representing the categories of one variable and columns representing the categories of another. When a dependent variable can be identified, it is customary to make this the row variable and to treat the independent variable as the column variable, although this arbitrary convention is often broken for formatting considerations.

Let us first consider the cross-tabulation of the two nominal-scale variables from the 2006 GSS shown in Table 15.4. The row variable consists of "attitude toward capital punishment" or, more precisely, whether the respondent favors or opposes the death penalty for persons convicted of murder. The column variable is "gender." Gender is the independent variable simply because, in one's lifetime, one's gender is determined earlier than one's attitude toward capital punishment. What sort of information does this table convey?

First, notice that the last column and the bottom row, each labeled "Total," show the total number of respondents with each single characteristic, for example, 1276 men. Because these four numbers (1945, 870, 1276, 1539) are along the right side and the bottom margin of the table, they are called **marginal frequencies**, or *marginals*. The row marginals (1945, 870) are the univariate frequency distribution for the variable "attitude toward capital punishment"; the column marginals (1276, 1539) are the univariate frequency distribution for the variable "gender." Also, the number at the lower right-hand corner (2815) is N, the total sample size excluding missing cases. N equals either the sum of the row or column marginals, or the sum of the four numbers (931 + 1014 + 345 + 525) in the body of the table.

TABLE 15.4. Attitude toward Capital Punishment by Gender, 2006 General Social Survey

*Attitude toward capital punishment**	*Gender*		
	MALE	FEMALE	TOTAL
Favor	931	1014	1945
Oppose	345	525	870
Total	1276	1539	2815

*Question: "Do you favor or oppose the death penalty for persons convicted of murder?"

The body of the table where the categories of the two variables intersect contains the bivariate frequency distribution. Each intersection is called a *cell* and the number in each cell is called a **cell frequency**. Cell frequencies in a *bi*variate table indicate the numbers of cases with each possible combination of *two* characteristics; for example, there were 931 *men* who *favored capital punishment.* Because Table 15.4 has two rows and two columns, it is referred to as a *2 × 2 table.*

Now that we know the meaning of the numbers in a cross-tabulation, how do we analyze these numbers to assess the relationship between the variables? What comparisons should we make? With gender as the independent variable in Table 15.4, the relevant substantive questions are as follows: Does gender influence attitude toward capital punishment? If so, how much influence does it have? If a relationship exists, then a change in gender should produce a change of "favor"/"oppose" responses in the distribution on the dependent variable. Either men will be more likely to favor and less likely to oppose than women, or women will be more likely to favor and less likely to oppose than men. Notice, however, that when we compare cell frequencies there are more women than men who both favor (1014 versus 931) and oppose (525 versus 345).

Obviously, there is a problem here. Although we are comparing the proper cells, the comparison is invalid because the cell frequencies for men and women are based on different total frequencies. To be valid, the cell numbers compared must be expressed as parts of the same total. This is accomplished by creating separate percentage distributions for men and women, thereby converting each total to 100 percent. The result is a bivariate percentage distribution, presented as Table 15.5. Now when we compare responses across gender, we see clearly that men are more likely to favor capital punishment by a percentage of 73.0 to 65.9 and, conversely, less likely to oppose (27.0 percent to 34.1 percent).

A bivariate percentage distribution enables one to compare the distribution of one variable across the categories of the other. In Table 15.5, we created such a distribution by percentaging *down* so that the column totals, corresponding to the categories of the independent variable, equaled 100 percent. The rule that we followed in deriving this table is to *compute percentages in the direction of the independent variable* (i.e., based on the categories of the independent variable). If gender had been the row variable and attitude toward capital punishment the column variable, we would have run the percentages in the other direction—across rather than down.

TABLE 15.5. Attitude toward Capital Punishment by Gender (%), 2006 General Social Survey

Attitude toward capital punishment	*Gender*	
	MALE	FEMALE
Favor	73.0%	65.9%
Oppose	27.0	34.1
Total	100.0%	100.0%
(*N*)	(1276)	(1539)

To interpret the relationship in Table 15.5, we compared percentages by reading *across* the table. In so doing, we followed a second rule: *Make comparisons in the opposite direction from the way percentages are run.* Having percentaged down, we compared across; had we percentaged across, we would have compared down.

KEY POINT

The rule for interpreting a cross-tabulation is to compare percentages across the variable categories on which the percentages are based; therefore, percentage across, read down or, conversely, percentage down, read across.

These are extremely important rules to follow, because cross-tabulations may be percentaged in either direction and are easily misinterpreted.[7] If we percentage Table 15.4 in the direction of the attitude variable, the results would indicate whether those who favor capital punishment are more likely to be men than those who oppose, which is an odd and unlikely research question.

Broh might have done bivariate analyses to see if there were gender differences on her key variables. Table 15.6 presents the percentage distribution of student reports of school discussions with their parents by gender.[8] What does this table reveal about the relationship between these variables? It clearly shows that in this sample of students girls report more school discussions with their parents than do boys. But does this necessarily mean that the relationship holds for the population from which the sample was drawn?

As you read across Table 15.6, you see that there is a small difference of 9.9 (46.4 – 36.5) in the percentage of girls as opposed to boys who report three or more school discussions with their parents. This "percentage difference" indicates that a relationship exists for these data; if there were no difference between the percentages, we would conclude that no relationship exists. Remember, however, that these are *sample* data. The important question is not whether a relationship exists in these data; rather, do the observed cell frequencies reveal a true relationship between the variables in the *population*, or are they simply the result of sampling and other random error?

TABLE 15.6. Discuss Programs at School* with Parents by Gender (%), NELS:88 Base-Year

Discuss selecting courses or programs at school with parents	*Gender*	
	MALE	FEMALE
Not at all	14.8%	8.7%
Once or twice	48.7	44.9
Three or more times	36.5	46.4
Total	100.0%	100.0%
(*N*)	(5661)	(5819)

*Question 36a: "Since the beginning of the school year, how often have you discussed the following with either or both of your parents or guardians?—Selecting courses or programs at school."

The latter judgment is made by means of *tests of statistical significance.* For cross-tabulations, the most commonly used statistic is the **chi-square** (*or* χ^2) **test for independence.** The chi-square test is based on a comparison of the observed cell frequencies with the cell frequencies one would expect if there were no relationship between the variables. Table 15.7 shows the expected cell frequencies, assuming no relationship, and the derived bivariate percentage distribution. Notice that the cell percentages in Table 15.7 (reading across) are the same as the marginals; this indicates that knowing whether a respondent is male or female is of no help in predicting school discussions with parents, precisely the meaning of "no relationship" between variables. The larger the differences between the actual cell frequencies and those expected assuming no relationship, the larger the value of chi-square and the more likely that the relationship exists in the population. Chi-square values for the data in Tables 15.5 and 15.6 are both statistically significant.[9] This suggests that in the American population, women are less likely than men to favor capital punishment (Table 15.5), and that high school girls are more likely than high school boys to discuss school courses or programs with their parents (Table 15.6).

Knowing that these relationships exist in the population, however, does not tell us how much effect the independent variable has on the dependent variable. It is possible for a relationship to exist when changes in one variable correspond only slightly to changes in the other. The degree of this correspondence is a second measurable property of bivariate distributions. In a 2 × 2 table, the percentage difference provides one indicator, albeit a poor one, of the strength of the relationship: The larger the difference, the stronger the relationship. However, researchers prefer to use one of several other statistics to measure the size of this effect. These **measures of association** are standardized to vary between 0 (no association) and plus or minus 1.0 (perfect association). One such measure, commonly used for 2 × 2 tables, is Yule's *Q*, which equals .17 for the data in Table 15.4.[10] Although the choice of labels is somewhat arbitrary, the magnitude of *Q* suggests a "low" association between gender and attitude toward capital punishment (see Davis, 1971:49). (For variables with nominal categories, the sign, + or − does not reveal anything meaningful about the nature of the relationship.)

TABLE 15.7. Discuss Programs at School with Parents by Gender, Assuming No Relationship

Discuss selecting courses or programs at school with parents	*Frequencies*			*Percentages*		
	MALE	FEMALE	TOTAL	MALE	FEMALE	TOTAL
Not at all	663	682	1345	11.7%	11.7%	11.7%
Once or twice	2647	2720	5367	46.8	46.8	46.8
Three or more times	2351	2417	4768	41.5	41.5	41.5
Total	5661	5819	11,480	100.0%	100.0%	100.0%
(*N*)				(5661)	(5819)	(11,480)

A third property—the direction of the relationship—may be measured when the categories of both variables can be ordered, that is, when both have at least ordinal-level measurement. *Direction* refers to the tendency for increases in the values of one variable to be associated with systematic increases or decreases in the values of another variable. Both variables may change in the same direction (a positive relationship) or in opposite directions (a negative relationship). In a positive relationship, lower values of one variable tend to be associated with lower values of the other variable, and higher values of one variable tend to go along with higher values of the other. The categories of the variables of education and income, for example, can be ordered from low to high. When we examine the association between these variables we expect it to be positive, with less educated persons tending to have lower incomes and persons of higher education tending to have higher incomes.

Table 15.8 shows a positive relationship between two 2006 GSS ordinal variables, level of education (highest degree obtained) and self-reported happiness. Carefully examine this table before reading further. We will consider happiness and highest degree as dependent and independent variables, respectively, since educational attainment occurs earlier in time and is a much more permanent condition than the expression of happiness given during the interview. Notice that, as in the previous examples, Table 15.8 is percentaged down for each category of the independent variable, highest degree; thus, comparisons should be made across. The percentage of persons "not too happy" (first row) generally falls with increasing education: 17.3 to 12.7 to 12.9 to 5.6 to 8.0 percent. Similarly, the percentage claiming that they are "very happy" (third row) consistently rises as educational attainment increases: 23.7 to 30.1 to 30.1 to 41.4 to 44.2 percent. It is this sort of pattern (greater happiness associated with higher education) that suggests a clearly positive relationship.

In a negative (inverse) relationship, there is a tendency for *lower* values of one variable to be associated with *higher* values of the other variable. Table 15.9 reveals such a relationship between education and the number of hours of television watched on an average day: As education increases, television viewing time decreases. (There are many possible explanations here: Perhaps higher-income persons

TABLE 15.8. General Happiness by Highest Degree Received, 2006 General Social Survey

	Highest degree				
*Happiness**	LESS THAN HIGH SCHOOL	HIGH SCHOOL	ASSOCIATE/JUNIOR COLLEGE	BACHELOR DEGREE	GRADUATE DEGREE
Not too happy	17.3%	12.7%	12.9%	5.6%	8.0%
Pretty happy	59.0	57.2	57.0	53.0	47.8
Very happy	23.7	30.1	30.1	41.4	44.2
Total	100.0%	100.0%	100.0%	100.0%	100.0%
(*N*)	(433)	(1501)	(271)	(531)	(255)

*Question: "Taken all together, how would you say things are these days—would you say that you are very happy, pretty happy, or not too happy?"

TABLE 15.9. Number of Hours per Day Watching TV by Highest Degree Received, 2006 General Social Survey

Number of *TV hours*	*Highest degree*				
	LESS THAN HIGH SCHOOL	HIGH SCHOOL	ASSOCIATE/JUNIOR COLLEGE	BACHELOR DEGREE	GRADUATE DEGREE
0–1	18.7%	24.2%	32.8%	33.3%	41.9%
2–3	41.5	46.4	44.1	51.1	47.4
4 or more	39.8	29.5	23.1	15.7	10.7
Total	100.0%	100.0%	100.0%	100.0%	100.0%
(*N*)	(262)	(996)	(193)	(3620)	(172)

are more apt to be exposed to and can better afford other, more expensive forms of entertainment or perhaps higher-educated people have less leisure time or maybe they read more.)

The statistical significance of a relationship between two ordinal-scale variables also may be tested with the chi-square statistic. Ordinal measures of the strength of association, while differing in name from nominal measures of association, are similar in concept. The magnitude of such statistics, ignoring the sign, indicates the strength of the relationship; and the sign (+ or −) indicates the direction of the association (positive or negative). One such statistic, gamma, equals .21 for the data in Table 15.8 and −.27 for the data in Table 15.9.[11] Thus, there is a low positive association between happiness and educational attainment and a moderate negative association between television hours and years of schooling.

Armed with the knowledge of how to interpret cross-tabulations, you are ready to test a bivariate hypothesis. We will illustrate hypothesis testing first with the results from a split-ballot wording experiment from the 2006 GSS. The GSS uses a series of questions about spending priorities in key policy areas to track trends in public support for various government spending programs. Respondents are asked whether they think "we're spending too much money on it, too little money, or about the right amount" on various labeled problems (crime, welfare, national defense, education, etc.). Since the "welfare" label may convey unfavorable connotations, it was paired with an alternative "assistance to the poor" label in one of several wording experiments. Respondents were randomly assigned to either the "welfare" or to the "assistance to the poor" question version. Table 15.10 reveals that the wording manipulation (independent variable) produced dramatic differences in support for government spending (dependent variable). Only 24.4 percent in the "welfare" condition thought too little money was being spent compared to 68.3 percent in the "assistance to the poor" condition. Similarly, the percentage responding that government spending was too much increased from 7.6 percent for the "poor" to 37.2

KEY POINT

Measures of association indicate how strongly variables are related; *tests of statistical significance* indicate the likelihood that the relationship occurred by chance.

TABLE 15.10. Split-Ballot Wording Experiment, 2006 General Social Survey

	Experimental wording	
Response	"ASSISTANCE TO THE POOR"	"WELFARE"
Too little	68.3%	24.4%
About right	21.9	35.5
Too much	7.6	37.2
Don't know	2.2	2.9
Total	100.0%	100.0%
(*N*)	(1529)	(1464)

*Question: "We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount. Are we spending too much money, too little money, or about the right amount on . . . [assistance to the poor/welfare]?"

percent for the "welfare" label. Since the split-ballot was a true experimental design, the only rival explanation for the wording results is that they were created by prior subject differences uncontrolled by random assignment. This is ruled out by a statistically significant chi-square test result.[12]

In a true experiment, the analysis usually begins and ends with an empirical examination of the hypothesized relationship between the dependent variable and the independent variable (experimental treatment). In surveys and other nonexperimental designs, the bivariate association between the dependent and independent variables is only the first step toward a multivariate analysis as extraneous variables may pose rival explanations for the results. Next, we present two nonexperimental examples of preliminary hypothesis testing, and we introduce statistical techniques that are the foundation for the advanced multivariate techniques discussed in chapter 16.

So far we have restricted ourselves to variables having only two to five categories and to tables with four to fifteen cells. This is not unusual because most cross-tabulation analyses in social research are limited to variables with relatively few categories. There are three important reasons for this. First, the size of the table increases geometrically as the number of categories for each variable increases. And the larger the table, the more difficult it is to discern the pattern of the relationship, which can be much more complex than the positive or negative relationships we have described. Second, as we pointed out in our discussion of sampling, the finer the breakdown of one's sample into various categories, the fewer cases there will be for any given breakdown (or cell of the table). Hence, larger tables may require impractically large samples for reliable assessments. Finally, variables with a relatively large number of categories either constitute or tend to approximate interval-scale measurement. With interval-scale variables, we can use a more precise and more powerful form of statistical analysis known as "correlation" and "regression."

Interval- and Ratio-Scale Variables

In chapter 4 we showed how relationships between two quantitative variables are depicted by plotting the values of each variable in a graphic coordinate system. In conjunction with this form of presentation, social researchers use a statistical method called **regression analysis** to analyze the effect of one interval/ratio variable on another. This is done by finding the mathematical equation that most closely describes the data.

We will use previously described variables from the country data set (Figure 15.1) to explore the hypothesis that countries with low rates of female literacy tend to have high rates of fertility. Let us begin our analysis by looking at a **scatterplot** of total fertility rates and female literacy rates for 178 countries (Figure 15.4). Each plot or point in the graph represents the values of one of the 178 countries on two variables. With the vertical axis as our reference, we can read the value of the dependent variable (total fertility rate); and with the horizontal axis as our reference, we can read the value of the independent variable (female literacy rate). The country of Niger, for example, has the highest fertility rate (7.1) and the lowest female literacy rate (10%). The lowest total fertility rates belong to Hong Kong (1.0) and Georgia (1.1), which have high female literacy rates (90% and 98%, respectively).

The scatterplot gives the researcher a rough sense of the form of the relationship: whether it is positive or negative and whether it is best characterized with a

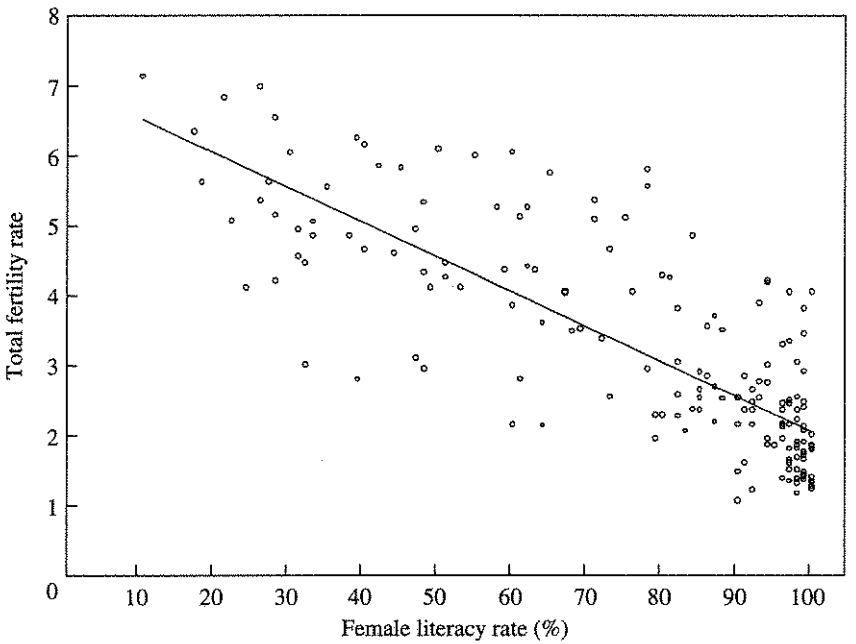

FIGURE 15.4. Scatterplot of total fertility rate by female literacy, 178 countries. Missing female literacy values for fifteen high-literacy countries were replaced with the country's total literacy rates. Countries with missing values for total literacy and/or fertility were excluded from the analysis. *Sources*: *The World Factbook* (http://www.cia.gov/cia/publications/factbook/index.html) and World Bank Developmental Data (http://devdata.worldbank.org/dataonline/).

straight or a curved line. This is crucial information because regression analysis assumes that the data have a particular form. If a straight line provides the best fit with the data, one should do linear regression; if a curve provides the best fit, one should use special techniques for fitting curvilinear relationships (which are beyond the scope of this book). As a first step in our analysis, we will examine the scatterplot to make a judgment about form. The overall form of the data in Figure 15.4 is somewhat difficult to discern; however, since the relationship does not appear to be sharply curvilinear, we can assume that a straight line offers as good a fit as a curved line. The trend of the data also suggests that fertility decreases as literacy increases.

Having decided to fit a straight line to the data, and therefore to do *linear* regression analysis, we need to know two things: (1) the mathematical equation for a straight line and (2) the criterion for selecting a line to represent the data.

The general form of the equation for a straight line is $Y = a + bX$, where Y is the predicted value of the dependent variable and X is the corresponding value of the independent variable. Thus, an equation for a straight line relating female fertility and literacy is

$$\text{Total fertility rate} = a + b\ (\text{Female literacy rate})$$

The value a, called the **Y-intercept**, is the point where the line crosses the vertical axis (where literacy = 0). The value b, called the **slope** or **regression coefficient**, indicates how much Y increases (or decreases) for every change of one unit in X—in other words, how much increase (or decrease) occurs in the total fertility rate (number of children) for every change of 1 percent in female literacy. To get the line of best fit, then, we could simply draw a line on the scatterplot that seems to reflect best the trend in the data and then determine the values of a and b from the graph. Of course, there are many lines that we could draw—a and b can take on an infinite number of values. How, then, do we know when we have obtained the best fit?

Regression analysis uses the method of least squares as the criterion for selecting the line that best describes the data. According to this method, the best-fitting line minimizes the sum of the squared vertical distances from the data points to the line. We have drawn the **regression line,** also called the **"least squares line,"** on the scatterplot. Now imagine a dashed line showing the vertical distance, as measured by the fertility number of children, between a specific data point, say Niger, and the regression line. The regression line represents the equation for predicting Y from X; the vertical distances between data points and this line represent prediction errors (also called **residuals**). Thus, by finding the line that minimizes the sum of the squared distances from it we are, in effect, finding the best linear predictor of the fertility rate from knowledge of a country's rate of female literacy.

The precise equation generated by the method of least squares can be found via a mathematical formula with the aid of a computer program. When we applied this formula to the data in Figure 15.4, we got the following equation:

$$\text{Total fertility rate} = 6.98 - .05\ (\text{Female literacy rate})$$

In other words, women in a hypothetical country with 0 percent literacy would be predicted to average about seven children during their lifetime; and, for every ten-

unit increase in literacy (10 percent), a decrease of about a half a child (.50) is expected in birth rates. The regression equation gives the best linear prediction of the dependent variable based on the data at hand. Niger, for example, has a predicted fertility rate of 6.48 ($6.98 - .05 \times 10$), which is .62 less (the residual) than its current 7.1 fertility rate. Can you spot the largest residuals in Figure 15.4? They belong to the Republic of Congo (5.76 fertility, 78% literacy) and Bangladesh (2.95 fertility, 32% literacy).

The strength of the association between two variables measured at the interval/ratio level is frequently measured by the **correlation coefficient** (symbolized as r), which may vary between -1 and $+1$. The sign of the coefficient, which is always the same as the sign of the regression coefficient, indicates the direction of the relationship. The magnitude of its value depends on two factors: (1) the steepness of the regression line and (2) the variation or scatter of the data points around this line. If the line is not very steep, so that it is nearly parallel to the X-axis, then we might as well predict the same value of Y for every unit change in X as there is very little change in our prediction (as indicated by b in the equation) for every unit change in the independent variable. By the same token, the greater the spread of values about the regression line (regardless of the steepness of the slope), the less accurate are predictions based on the linear regression. The scatterplot for the regression of fertility on literacy shows that the line is fairly steep in relation to the horizontal axis and that the data points (countries) cluster fairly close to the line. Not surprisingly, therefore, the correlation coefficient indicates a strong negative association of $-.80$. Statistics also exist for testing whether the correlation coefficient and the regression coefficient are significantly different from zero. These may be found in most statistics textbooks. Both of these coefficients are significant ($p < .001$) for the data in Figure 15.4.

The strong negative association between countries' fertility and female literacy rates by itself, however, does not tell us whether there is a causal connection between the two. To establish this, we need to show that the direction of influence is from literacy to fertility (the presumed independent and dependent variables, respectively) and that the association between the variables is not spurious. This, in turn, requires a country-level theoretical model of fertility determinants, along with a multivariate analysis to rule out rival explanations of the negative association. At the country (aggregate data) level, both high fertility and low female literacy may reflect discrimination against women in traditional societies, as well as low income levels, high infant mortality rates, government opposition to family planning, poor health care, and so forth.

KEY POINT

Regression coefficients indicate the direction and amount of change in the dependent variable for each change of one unit in the independent variable. *Correlation coefficients* add information about the strength of the relationship or how well a linear regression predicts the dependent variable.

Our second example of regression analysis is from Broh's (2002) study of extracurricular activities and academic achievement. In a preliminary analysis, she explored the hypothesis that playing high school interscholastic sports (independent variable) improves students' math grades (dependent variable). Since the NELS longitudinal (panel) study collected information from 8th graders in 1988 (baseline) with follow-ups when most were in the 10th (1990) and 12th (1992) grades, she ex-

amined if the interscholastic sports participants during both the 10th and 12th grades ("athletes") had higher 12th-grade math grades than those who did not participate in the 10th and 12th grades ("nonathletes").

Regression analysis can be applied to independent variables with only two categories, such as Broh's comparison of "athletes" and "nonathletes," by creating what is called a **dummy variable**. This is a variable that has been recoded so that one of its categories has a value of 1 and the other category has a value of 0. Dummy coding enables the researcher to manipulate the variable numerically and to use certain kinds of statistical analysis that would not be possible otherwise. We could, for example, code women as 1 and men as 0 and proceed to regress an interval-/ratio-level variable, such as number of television hours watched, on gender. The horizontal axis would have only two values, 0 and 1, and the regression coefficient would indicate the difference between men and women in the mean number of television hours watched. Broh used dummy coding to measure interscholastic sports performance during both the 10th and 12th grades (1 = participated in both years, 0 = did not participate in both years).

Because the scaling of some NELS measures differed in baseline and follow-up years, Broh put them on equal footing by re-expressing them in standard score form (mean = 0, standard deviation = 1).[13] A student with math grades right at the mean for all NELS students in 1990 and 1.5 standard deviation below the mean in 1992, for example, would have standardized scores of 0 and −1.5, respectively. When Broh regressed interscholastic sports participation on 12th-grade (1992) math grades for 10,379 students, she got the following equation:

$$1992 \text{ math grades} = .005 + .230 \text{ (athlete in 10th \& 12th grade)}$$

In other words, the regression equation estimates that athletes in both the 10th and 12th grades will have math grades in 1992 about one-quarter (.230) of a standard deviation higher than students not playing interscholastic sports.[14] Although the higher math grade for interscholastic athletes was statistically significant ($p < .001$), the observed association between playing sports and math grades may be spurious. Smarter students, for example, might be more likely than their peers to play interscholastic sports and to do well in math. In the next chapter we describe how Broh used multivariate statistical techniques to test for spuriousness and to model theorized causal links between high school sports participation and educational achievement.

Summary

Data analysis involves an iterative comparison of theory and data. Data analysis begins and ends with the researcher's hypotheses or theoretical model. Such models guide the collection and analysis of data, indicating what variables should be measured and statistically controlled. The analyses, in turn, suggest new theoretical formulations and new directions for research.

Before quantitative analysis can begin, the data are processed and put in computer-readable form. In survey research, this entails four steps: editing, coding,

data entry, and cleaning. Some of these procedures occur during data collection in computer-assisted surveys. Editing is designed to ensure that the data to be entered into the computer are as complete, error-free, and readable as possible. Coding consists of assigning numbers to the categories of each variable. Most coding in survey research is done before data collection since answer codes for fixed-choice questions appear directly on the survey form. Codes for open-ended questions are developed after the data are collected.

After editing is complete, the data are entered into the computer and stored in a data file. Typically, a data set is organized as a matrix or spreadsheet, with observations as rows and variables as columns. After entry, the data are cleaned for errors in coding and transmission to the computer. This is a multistep process, sometimes beginning with a verification procedure whereby all data are entered into the computer a second time and each dual entry is checked for noncomparable codes. Wild-code checking consists of checking for "illegal" codes among the values recorded for each variable. Consistency checking consists of checking for reasonable patterns of responses, such as whether contingency questions are answered only by those for whom the questions are intended.

With the data cleaned, the researcher usually prepares a codebook, which gives the answers that correspond to each numerical code, the location of each variable in the data file, and coding and decision rules. At this point, the researcher is ready to inspect and modify the data for the planned analysis. The goal of inspection is to get a clear picture of the data by examining each variable singly (univariate analysis). Data modifications include changing one or more variable codes, rearranging the numerical order of variable codes, collapsing variable categories, imputing estimated values for missing data, and adding together the codes for several variables to create an index or scale.

The primary functions of statistics are descriptive and inferential. Descriptive statistics are designed to summarize the data at hand; inferential statistics indicate the extent to which one may generalize beyond the data at hand. In univariate analysis the categories or values of each variable first are organized into frequency and percentage distributions. If the data constitute interval-level measurement, the researcher will also compute statistics that define various properties of the distribution. Statistical measures of central tendency reveal various points of concentration: the most typical value (mode), the middle value (median), and the average (mean). Common measures of dispersion are the difference between the lowest and highest values (range) and an index of the spread of the values around the mean (standard deviation). Distributions also may be described in terms of their shape, such as their skewness.

Bivariate analysis examines the nature of the relationship between two variables. For relationships involving exclusively nominal- or ordinal-scale variables, such analysis begins with the construction of cross-tabulations, with table cells containing the bivariate distribution. The rule for percentaging cross-tabulations is to percentage in the direction of the independent variable; the rule for reading such tables is to make comparisons in the direction opposite the way the percentages are run. For relationships involving interval- or ratio-scale variables, the data are plotted in a scatterplot and characterized in terms of a mathematical function. Linear regression analysis

identifies the straight-line function that provides the best fit with the data by virtue of minimizing the sum of the squared deviations from the line. The slope of the line reveals the predicted change in the dependent variable per unit change in the independent variable. Nominal-/ordinal-scale variables may be included in regression analyses as independent variables by dummy-coding categories 0 and 1, in which case the slope indicates the mean difference between categories on the dependent variable.

Regardless of measurement level, bivariate analysis assesses both the likelihood that the observed relationship is the product of random error and the strength of the association between the variables. The first task usually is accomplished for cross-tabulations by means of the chi-square test for independence. Statistical measures of association generally vary from +1 to −1, with the sign indicating the direction of the relationship and the magnitude indicating the strength of the association.

Bivariate analysis may be used to test simple two-variable hypotheses, which is the final analysis step in true experiments since an adequate design effectively controls extraneous variables as rival explanations for the results. But in non-experimental designs, extraneous variables may pose serious rival explanations that require statistical control in a multivariate model, which we discuss in the next chapter.

Key Terms

- data processing
- data modification
- editing
- coding
- data cleaning
  - wild-code checking
  - consistency checking
- data matrix
- codebook
- descriptive statistics
- inferential statistics
- univariate analysis
  - frequency distribution
  - percentage distribution
  - mean
  - median
  - mode
  - range
  - standard deviation
  - frequency/percentage polygon
  - normal distribution
  - outliers
- listwise deletion
- imputation
- bivariate analysis
  - marginal frequencies
  - cell frequencies
  - chi-square test for independence
  - measures of association
  - regression analysis
    - scatterplot
    - Y-intercept
    - slope/regression coefficient
    - regression (least squares) line
    - residuals
    - correlation coefficient
    - dummy variable

Exercises

1. To become familiar with data documentation, coding, and data modification, access the GSS Codebook online (http://www.norc.org/GSS+Website/Codebook/). Note that the codebook has two indexes, an index to the data set with

mnemonic labels and question numbers and a subject index to the questions (Appendix V). Using either index, answer the following questions.

a. Find the GSS questions on respondents' labor force status, number of siblings, attendance at religious services, knowledge of science, alcohol use, and job satisfaction.
b. The codebook indicates if questions were recoded, and Appendix D describes the recodes. Go to Appendix D. What question is asked in the GSS to measure age? According to Appendix D, why is this question asked and how are responses to the question recoded? Now find the variable MOBILE16 in the codebook. What does it measure? What are the response categories, and how were they created?
c. The 2000 GSS asked two open-ended questions on the meaning of freedom. Find these in the codebook. How many coding categories were created for the first freedom question: When you think about freedom, what comes to mind? How was this question coded when a respondent gave more than one answer?
d. The 2002 and 2004 GSS contained a module on altruism which included seven questions that measured empathy. Find "empathy" in the subject index under the heading "Feelings." Examine these questions in the codebook and decide how you will combine them to create an index.

2. This exercise asks you to test a bivariate hypothesis using GSS data. First, formulate a hypothesis by selecting one independent variable and one dependent variable from the following lists.

Independent Variables	Mnemonic Label
Gender (male/female)	SEX
Race (white/black)	RACE(1,2)
Level of education	DEGREE
Dependent Variables	**Mnemonic Label**
Support for legalization of marijuana (should/should not be legal)	GRASS
Attitude toward homosexuality (sexual relations between adults of same sex is always wrong or not wrong at all)	HOMOSEX(1,4)
Whether or not one has seen an X-rated movie	XMOVIE

Now go to the textbook Web site (college.holycross.edu/projects/approaches5). Select chapter 15, then "**Web Resources**," and then scroll down and click on "**Current User Interface**." This Web site contains a program for the quick and easy analysis of GSS data. To test your hypothesis, you will need to "run" a crosstabulation as follows: To the right of the page, below "**SDA Frequencies/Crosstabulation Program**," enter the mnemonic for your dependent variable as your Row variable; enter the mnemonic for your independent variable as your Column variable; then enter "YEAR(2006)" as your Selection Filter(s); select "Column" for Percentaging, "Statistics," and "Question text" under "**Table Options**"; finally, click on "Run the table" in the shaded box to perform the analysis.

a. Do these data support your hypothesis? Briefly explain.
b. Is the value of chi-square significant? What is the value of gamma? What do these statistics tell you about the relationship?

3. Consider the following values of Yule's *Q* for three different sets of variables: −.82, +.04, +.35. Which association is strongest? Which association is negligible?

4. Dummy-code the following variables: gender, attitude toward capital punishment (favor/oppose) race (white/black, treating "other" as missing). What does the mean value of a dummy variable indicate?

5. Consider the following equation based on 2006 GSS data:

Respondent's years of education = 9.56 + .359 (Father's years of education)

Describe the predicted relationship in this equation. How much change in respondent's education is associated with each increase of 1 year in father's education? What are the predicted years of education of a respondent whose father has completed 16 years of education? (This will require some calculation.)

Notes

1. There are other database options for organizing complex data sets. In contextual analysis, a single data file may contain, for example, both student-level and school-level variables. Or there could be a hierarchical arrangement of separate data files (e.g., neighborhoods, schools, classrooms, and students) with links identifying the students in a particular classroom, the classrooms in a particular school, and the schools in a particular neighborhood. In one social network design, the social actors are listed in both the rows and the columns of a data matrix, with the cell entries representing relationships or connections among pairs of actors.

2. The distinction between descriptive and inferential statistics should not be confused with the scientific goals of description and explanation. Both forms of statistical analysis may be used to accomplish both goals. The first purpose of inferential statistics identified here is description, and the second purpose is explanation. Similarly, descriptive statistics may provide a summary measure of some characteristic of the sample data (description) or may indicate the strength of a hypothesized association between two variables (explanation).

3. Chance may enter into theoretical explanations, which has important implications for popular statistical techniques (Berk, 1983).

4. Social scientists today rarely write their own programs of instructions to the computer. Instead, they rely on "canned" or "packaged" programs stored in the computer and specially written to perform statistical analyses of social science data. To use these programs, you must (1) give the computer a command that accesses the particular package; (2) using the rules for that package, describe the format and location of the data file; and (3) indicate the specific analyses you want done. Thorough discussions of these procedures can be found in the manuals for SPSS, Stata, SAS, or any other packages available for your computer.

5. Allison (2002) assumes that the chance of missing data on one variable is independent of the chance of missing data on any other variable.

6. See GSS Codebook (http://www.norc.org/GSS+Website/Codebook) Appendices F (Occupational Classification Distributions) and G (Prestige Scores).

7. Sometimes these rules are broken for descriptive reasons (e.g., to ascertain the educational background of different income groups) or because the data are unrepresentative of the population with respect to the dependent variable (see Zeisel, 1968:30–36).

8. Statistical analyses ordinarily require adjustments for complex sampling designs such as those used in the GSS and NELS. The NELS, for example, used disproportionate stratified cluster sampling, with students clustered by schools. For pedagogical reasons, Tables 15.1, 15.2, 15.6, and 15.7 are based on unweighted frequencies; all other tables are based on more appropriate weighted frequencies.

9. For Table 15.5, Pearson's chi-square statistic = 16.36, 1 df ($p < .001$); for Table 15.6, chi-square = 167.78, 2 df ($p < .001$).

10. For an explanation of Yule's Q and other popular measures of association, see chapter 5 in Bohrnstedt and Knoke (1994).

11. See Bohrnstedt and Knoke (1994:168–75) for a discussion of gamma.

12. For Table 15.10, excluding the "don't know" responses, chi-square = 651.75, df = 2 ($p < .001$).

13. Variables are standardized by subtracting the variable mean from each original score and then dividing by the standard deviation of the original scores.

14. The equation for athletes is .005 + .230(1) = .235; for nonathletes, .005 + .230(0) = .005.