CHAPTER 1

GETTING STARTED WITH STATISTICAL ANALYSIS

Where Do I Obtain Data, and How Do I Prepare Data for Statistical Analysis?

This is a book about statistics. It is not a book about research design. Nonetheless, the practice of statistical analysis is intricately related to issues of research methodology. These issues arise long before data can be massaged, manipulated, interpreted, and transformed into meaningful conclusions drawn from empirical research studies. You have probably heard the maxim "Garbage in, garbage out," in reference to the use of statistics. The implication is that without thoughtfully considering the source and quality of your data, the most sophisticated statistical analysis available won't be of much help.

In the first edition of this book, we assumed that readers would already have data available and would only need guidance through the steps of data coding, data entry, and data cleaning. Since that time, however, we have become more conscious of a prior question our students and colleagues often ask, "Where can I obtain some (good) data?" One rather glib response is "Go collect your own." However, collecting one's own data through the customary process of designing a study, administering a data collection instrument, coding the data, and entering the data into a computer, although within most social and behavioral science disciplines the predominant option, is only one approach. And it may not always be the best option. In some instances, it may be absolutely the worst option. Thus, we begin this new edition with a brief discussion of the available options for acquiring data, some of which require no data entry steps whatsoever.

WHERE DATA COME FROM

Primary Data Analysis

Researchers often make the distinction between "primary" data analysis and "secondary" data analysis. Primary data analysis is grounded in methodological processes that make use of experimentation, survey, or field methods that may require direct contact with the persons, organizations, or groups of interest to collect the data. There are also numerous ways that a researcher might obtain high-quality primary data without ever directly engaging either the data source or the data entry process. Information collected from these sources creates the units of analysis eventually comprising the database. In Chapter 5, we will consider issues of sampling that relate directly to the types and numbers of cases, respondents, or participants that might be chosen to provide the data for a suitable research study.

Data entry is time-consuming, error-prone, and not particularly fun. Therefore, any strategy that avoids the process, saves time, and facilitates accuracy is worth considering. In the traditional sequence, a questionnaire is completed by either the respondent (self-administered) or the interviewer. The questionnaire might be sent through the mail, completed via a face-to-face interview (survey interview), or completed over the phone. In some cases, the second step may be to "code" the questionnaire by writing a code for each response in spaces provided for this purpose on the questionnaire. However, these traditional procedures require a data entry step.

Are there methods that allow the data to be directly entered into a database? The answer is a clear yes. For example, phone surveys often use computer-aided telephone-interviewing (CATI) systems to directly enter data into a database. Even interviewer-administered surveys can include direct data entry via what is known as audio computer-assisted self-interviewing (ACASI). (See www.popcouncil.org/projects/246_ACASI.asp, for a detailed description of this technology.) We encourage you to pursue multiple possibilities for data entry beyond the traditional "by hand" methods.

It is also possible to design experimental studies and administer them on the Internet. By setting up the study properly, one can automate everything, from the presentation of stimulus materials to the recording, computation, and downloading of quantitative results. Commercial websites are available for this purpose, or you could consider hosting your own study. An early, informative guide to conducting behavioral research on the Internet is written by Michael Birnbaum (2000). A more recent overview is a chapter on conducting online research by

Reips and Krantz (2010). Another worthwhile resource is a report of the Board of Scientific Affairs Advisory Group on the Conduct of Research on the Internet that identifies the benefits and challenges of conducting research via the Internet (Kraut et al., 2003). Don Dillman's *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method* (2008) also offers a wealth of information on good survey design and how to use the Internet to collect survey data. A final recommended reference is an article titled "Researching Internet-Based Populations: Advantages and Disadvantages of Online Survey Research, Online Questionnaire-Authoring Software Packages, and Web Survey Services" (Wright, 2005). Table 1.1 offers an annotated selection of sites for Internet-based survey research.

Table 1.1 Online Services for Survey Design and Data Collection

Company Name/Product	*Site Address*
Active Web Survey	http://www.activewebsoftwares.com/activewebsurvey/
Apian Software	http://www.apian.com/
CreateSurvey	http://www.createsurvey.com/
EZSurvey	http://www.raosoft.com/ezsurvey/
FormSite	http://www.formsite.com/
InfoPoll	http://infopoll.com/live/surveys.dll/web
InstantSurvey	http://www.instantsurvey.com/
KeySurvey	http://www.keysurvey.com/
PollPro	http://www.pollpro.com/
SumQuest	http://download.cnet.com/SumQuest/3000-2064_4-10215637.html
SuperSurvey	http://www.supersurvey.com/
SurveyCrafter	http://www.surveycrafter.com/interim2/default.asp
SurveyMonkey	http://www.surveymonkey.com/
Vovici	http://www.vovici.com/
Zoomerang	http://www.zoomerang.com/

NOTE: The above list is not inclusive. The interested reader should examine the sites above as examples of available options for online survey research. For a review of online survey research options, see Wright (2005).

Now it is not only possible, but commonplace, to conduct primary analyses on the Internet, either by mounting your own survey online or by using a site such as Survey Monkey (www.surveymonkey.com) to do it for you. One noteworthy advantage of conducting survey research through a site such as Survey Monkey is that the data come to you in the format that you design, eliminating most of the hassles with secondary analysis described in the following and entirely eliminating the drudgery of the data entry process. These sites also support much of the work of formatting pleasing-looking surveys. With survey research conducted through an online service, you will not be required to enter data, and the costs of conducting mail, phone, or face-to-face based survey research will be eliminated. However, no site can make up for bad survey questions, so be sure to educate yourself about what makes a good question and how to organize a survey before jumping online.

Secondary Data Analysis

Secondary analysis refers to the use of data that are already available in one form or another. Widely available sources of secondary data include public or private documents and records, newspapers, magazines, blogs, video archives, and other mass media. However, the use of such existing materials requires an intermediate coding process to transform the content into data that can be analyzed statistically, essentially turning qualitative content, such as text or video, into quantitative content. Hence, the term *content analysis*. While content analysis is sometimes included within the realm of secondary data analysis, one might argue that content analysis is more akin to primary data analysis because the researcher often "samples" from such material and engages the data in a complex and lengthy transformative process.

True secondary analysis involves the use of an ever expanding and extensive collection of archival databases available on the Internet. These databases include wide-ranging data collected through federal and nonprofit grants, data collected by polling organizations such as Gallop and Roper, and data from the US Census. These data are often directly downloadable into statistical software packages such SAS or SPSS.

Secondary analyses of data downloaded from a web-based archive are becoming commonplace and involve the following steps.

1. *Locate the database.* This step usually involves searching one of many widely available data repositories. The goal is to find the best match possible between your interests and the available data. In Table 1.2 we list over 50 data archives in a wide variety of scientific disciplines. We suggest that you start

with the Inter-University Consortium for Political and Social Research (www.icpsr.umich.edu), the world's largest archive of social science data. The ICPSR describes itself in this way:

> An international consortium of about 700 academic institutions and research organizations, ICPSR provides leadership and training in data access, curation, and methods of analysis for the social science research community. ICPSR maintains a **data archive** of more than 500,000 [yes, that is 500,000] files of research in the social sciences. It hosts 16 specialized collections of data in education, aging, criminal justice, substance abuse, terrorism, and other fields.

From the home page you can click the "Find & Analyze Data" link, and you're on your way!

Despite the apparent advantage of using secondary data, don't forget that someone else's data may not contain the instrumentation or content you need to directly address your primary research questions. This results in either changing the questions to "fit" the available instrumentation or trying to create an instrument from parts of someone else's instruments. The problem is that these derived measures do not have the record of reliability and validity studies that support the use of an established measure. Thus, using established measures

(*Text continues on page 14.*)

Table 1.2 Data Archives and Libraries

Center of Demography and Ecology (University of Wisconsin–Madison)
http://www.ssc.wisc.edu/cde/
This research cooperative boasts "one of the country's finest collections of machine-readable data files in demography." Search the Data & Information Services Center (DISC) for information about data archives and user guides.

Consortium for International Earth Science Information Network (Columbia University)
http://www.ciesin.org/
Data on world population, environment, health, and geography. Includes several interactive systems to search for data.

Council for European Social Science Data Archives
http://www.nsd.uib.no/cessda/
Search for data available from archives around the world.

(Continued)

Table 1.2 (Continued)

Data & Information Services Center (University of Wisconsin–Madison)
http://www.disc.wisc.edu/
A collection of social science and cross-disciplinary datafiles. Some of the data collected here are available for purchase and/or download.

International Social Survey Programme
http://www.issp.org/
Cross-national collaboration on social science surveys in 34 countries. Served by the data archive ZentralarchivfürEmpirischeSozialforschung (Universitätzu Köln): http://www.gesis.org/das-institut/ (in German).

Inter-university Consortium for Political and Social Research
http://www.icpsr.umich.edu/
Maintains and provides access to a vast archive of social science data for research and instruction and offers training in quantitative methods to facilitate effective data use. Provides a searchable database of archival holdings as well as direct downloading of data for member institutions. Also cosponsors the following special topics archives:

Health and Medical Care Archive
http://www.icpsr.umich.edu/HMCA/

National Archive of Computerized Data on Aging
http://www.icpsr.umich.edu/NACDA/

National Archive of Criminal Justice Data
http://www.icpsr.umich.edu/NACJD/

Substance Abuse and Mental Health Data Archive
http://www.icpsr.umich.edu/SAMHDA/

National Data Archive of Child Abuse and Neglect
http://www.ndacan.cornell.edu/
Contains information regarding organization's mission, publications, and available datasets.

Odum Institute Data Archive (University of North Carolina)
http://www.irss.unc.edu/odum/
Public opinion data from the Louis Harris polls, Carolina and Southern Focus Polls, and the National Network of State Polls. Includes a searchable database to retrieve questions and frequencies. Selected data files also available for downloading.

Roper Center Public Opinion Archives (University of Connecticut)
http://www.ropercenter.uconn.edu/
Provides an extensive archive of opinion polls, including Gallup polls and many others.

Social Science & Government Data Library (University of California–Berkeley)
http://sunsite.berkeley.edu/GovData/info/
Provides interactive access to selected 1990 Census data files, including several Subject Summary Tape Files (SSTF), and is a mirror site for the Census Lookup system (1990 Summary Tape Files 1 and 3). Also provides FTP access to numerous Census data files, including 1970 Census Fifth Count data.

UK Data Archive (University of Essex)
http://www.data-archive.ac.uk/
Archives approximately 7,000 datasets in the social sciences.

US National Archives and Records Administration—Center for Electronic Records
http://www.archives.gov/research/
Information regarding electronic records, including numeric datafiles, generated by US government agencies and available for purchase through the National Archives and Records Administration.

Social and Political Data

American National Election Studies
http://www.electionstudies.org/
Conducts national surveys of the American electorate. The time-series of studies now spans five decades. This site provides information on the mission and procedures of the ANES and other documentation.

Federal Election Commission
http://www.fec.gov/
Offers downloadable data on campaign financing.

Gallup
http://www.gallup.com/
Public opinion data from Gallup, including some tables and statistics and articles from its newsletter and other reports.

General Social Survey (University of Chicago)
http://www3.norc.org/GSS+Website
Information regarding the biennial personal interview survey conducted by the National Opinion Research Center. Includes a search engine to search the codebook for relevant variables and an extract utility to subset data.

Living Standards Measurement Study (World Bank)
http://www.worldbank.org/LSMS/
Includes household surveys for numerous countries. Access conditions vary by country.

(Continued)

Table 1.2 (Continued)

Panel Study of Income Dynamics
http://psidonline.isr.umich.edu/
Information regarding the Panel Study of Income Dynamics, a longitudinal study of American families ongoing since 1968. Covers topics such as employment, income, wealth, housing, and health.

Uniform Crime Reports
http://fisher.lib.virginia.edu/collections/stats/crime/
Interactive system for retrieving county-level crime and arrest data.

United Kingdom Election Results
http://www.election.demon.co.uk/
Provides links to election results from British parliamentary elections since 1983.

Economic Data

Economics, Statistics, and Market Information System (US Department of Agriculture)
http://usda.mannlib.cornell.edu/
Publications and datasets about agriculture available from the statistical units of the USDA: Economic Research Service, National Agricultural Statistics Service, and World Agricultural Outlook Board.

Government Publications: Statistics (University of Minnesota)
https://www.lib.umn.edu/govpubs/statistics/
Links to selected statistical tables, publications, and indicators arranged by subject.

HUD USER (US Department of Housing and Urban Development)
http://www.huduser.org/
Provides data pertaining to housing needs, market conditions, and community development.

Government Statistical Agencies

Centers for Disease Control and Prevention, National Center for Health Statistics
http://www.cdc.gov/nchs/

Centers for Medicare & Medicaid Services
http://www.cms.gov/

FedStats
http://www.fedstats.gov/
Links to most US federal government sites providing access to statistical data.

Internal Revenue Service: Tax Stats (including Statistics of Income)
http://www.irs.gov/taxstats/

National Science Foundation, National Center for Science and Engineering Statistics
http://www.nsf.gov/statistics/

US Census Bureau
http://www.census.gov/

US Department of Agriculture, Economic Research Service
http://www.ers.usda.gov/

US Department of Commerce, Bureau of Economic Analysis
http://www.bea.gov/

US Department of Education, National Center for Education Statistics
http://nces.ed.gov/

US Department of Energy, Energy Information Administration
http://www.eia.gov/

US Department of Justice, Bureau of Justice Statistics
http://www.ojp.usdoj.gov/bjs/

US Department of Labor, Bureau of Labor Statistics
http://www.bls.gov/

US Department of Transportation, Bureau of Transportation Statistics
http://www.bts.gov/

Statistics Canada/Statistique Canada
http://www.statcan.gc.ca/

United Nations Statistics Division
http://unstats.un.org/unsd/default.htm

Other Website Directories

Data on the Net (University of California–San Diego)
http://3stages.org/idata/
Contains many links to sites with actual data along with lengthy descriptions of what can be found at each site. Can be searched by keyword.

Stephen S. Clark Library (University of Michigan)
http://www.lib.umich.edu/clark-library/
Predominantly links to government-produced data, arranged by subject.

when possible is recommended. Investing the time to locate the exact database that fits your questions is well worth the effort. In Table 1.3 we list some of the advantages and disadvantages of utilizing data from a web-based data archive.

Table 1.3 Advantages and Disadvantages of Using Secondary Data Sources

Advantages	*Disadvantages*
Data already exist. No data collection expenses.	No control over how data were collected.
Data already exist. No data entry expenses.	No control over how data were collected.
Sample may be drawn more scientifically. Encourages understanding of complexities of sampling and survey construction.	No control over sampling design. Sample may not be sufficiently narrow to reach desired population. Data may contain unknown sampling biases.
Data may be longitudinal: Avoids time and expense of a complex, lengthy data collection effort.	Longitudinal data may involve multiple complex data structures and documentation.
Sample may be larger than realistically collectable by a single researcher, encouraging statistical power.	No control over sampling design. Sample may not be sufficiently narrow to reach desired population. Data may contain unknown sampling biases.
You can benefit from the research from some of the top scholars in your field and for the most part be assured of quality data.	Data may contain unknown sampling biases.
Data may be easily accessible.	Data documentation may be inadequate, confusing, and complex. Multiple layers of permission may be required.
Data may be impossible to access in any other manner.	Data may be obsolete.
Use of multiple data sets may favor meta-analysis.	Data may not fit your research questions.
Very little of the available data may have been analyzed.	Data may not fit your research questions. No control over instrumentation.
Likely to be a link between database and many publications.	The question you wish to ask may have already been answered.
IRB issues likely to be minimal.	Secondary data can be used inappropriately. Data-sponsoring organization likely to impose additional requirements to protect participants' rights.

2. *Obtain the necessary passwords and permissions.* Satisfy any required Institutional Review Board agreements. For example, utilizing data from the ICPSR holdings may require that you belong to a member institution that has paid a fee for membership. To access some data from the National Data Archive on Child Abuse and Neglect (www.ndacan.cornell.edu), you may need to obtain a data license and/or satisfy additional requirements, such as agreeing to destroy the data after its intended use. While these steps are not insurmountable, be aware that they take time. Sometimes the process is more complex than simply finding a database and accomplishing an instant download.

3. *Master the download format or data extraction system.* This step can be easy or incredibly frustrating. Ideally the data will be available in a format that can be accessed with any statistics package, but this is not always the case. Extensive codebooks may accompany the data, and you may have to dig deep to find what you actually want. Do not assume that this will be easy. Despite the fact that most archive sites attempt to make data easy to locate and provide extensive documentation (some are more successful than others), the downloading process can be anything but simple. Be prepared to encounter difficulties. Not all archive sites use the same process for downloading or documenting data, and it might be necessary to learn several methods for doing so. In addition to feeling frustration when downloading and accessing a database with your own software, you may struggle with inadequate or inaccurate documentation. You must have access to a complete description of the sampling design, accurate and complete codebooks must be available, and, when all else fails, you must have a resource to contact for assistance when you encounter flaws or omissions in the available documentation.

4. *Access the downloaded data with statistical software.* Do *not* assume that just because the data were originally collected by high-powered researchers at a high-powered university that they will be perfect. The persons who actually enter and check the data may be far removed from these researchers. It is your job to study the data carefully and check for errors. There is no such thing as a "perfect" database, and once you assume responsibility for the data, you also assume responsibility for their accuracy.

In sum, there are great advantages to using archived data. Such data, particularly data derived from studies sponsored through federal grants, are likely to be of much better quality than those any independent researcher could collect. Large-scale research is often very expensive, and large-scale longitudinal research is both exorbitantly expensive and time prohibitive. Yet, such data are easily accessible through Internet-based data archives. For example, the National Longitudinal Survey of Youth 1997 (NLSY97) is part of the National Longitudinal Surveys (NLS) program, a set of surveys sponsored by the US Department of Labor,

Bureau of Labor Statistics (BLS). These surveys have gathered information at multiple points in time on the labor market experiences of diverse groups of men and women. Each of the NLS samples consists of several thousand individuals, some of whom have been surveyed over several decades (www.icpsr.umich.edu/NACJD/). These surveys are available from the National Archive of Criminal Justice Data, accessed through ICPSR. Clearly, access to data such as this should be considered as an option before one launches into an independent research project that will almost certainly pale in comparison to studies such as the NLS.

PREPARING DATA FOR ANALYSIS

This section concerns the preparation of data for analysis by a statistical program package such as SPSS® (Statistical Package for the Social Sciences), SAS® (Statistical Analysis System), or Stata®. To facilitate data entry, the data organization process must be planned in advance. There are two components of this process. The first concerns the coding of data in a manner that permits analysis by statistical software, and the second concerns thinking ahead about data entry. If the data are not properly prepared, the method used to enter them into a computer file (i.e., a database) will not fix the resulting problems.

There are several potential data entry alternatives, as data may be entered into database management programs (e.g., Microsoft Office Access) or spreadsheet programs (e.g., Microsoft Office Excel) or directly into statistical analysis programs through data entry options available with these programs. In addition, most major statistical programs, such as those mentioned above, can access data files created with a wide variety of software, including other statistical programs; however, it is still a good idea to investigate the data management capabilities of your statistical software before you enter your data, particularly if using a lesser-known program.

In sum, when organizing data that have been collected for analysis with statistical software, the user needs an understanding of how a data analysis program "reads" this information and reflects it back to the researcher in a meaningful way. The purpose of this section is to provide a guide for organizing data in a manner that permits a statistical program to access and analyze them. We begin with the basics: In the following paragraphs, we discuss cases, variables, values, codes, and our essential guidelines of data organization. Remember that there are exceptions to most of what we say below, and some would argue that there is a "better way." Our recommendations are designed to assist the novice and help all researchers avoid the most common mistakes.

Cases: How Do I Define a Unit of Analysis?

Measurement is defined as the process of assigning numbers to characteristics of persons, places, events, or things that are observable. In the process of observing and measuring, each person, place, event, or thing may be considered a primary unit of analysis and may be measured on a variety of characteristics. We call these basic units **cases.** For example, in laboratory experiments studying the growth of malignant cells, each animal treated with a different therapy represents one case. In social science surveys, each respondent may be considered one case. In the government's examination of the repayment of student loans, each loan becomes one case. We can see, therefore, that in each research project there will be a specific number of cases that will be analyzed. This is often referred to as the study's number of cases, sample size, or simply *N*.

How Do I Use Variables and Values to Describe a Case?

As we consider the measurement process, note that each object (i.e., each case) can be described with a variety of characteristics. Each of these characteristics is known as a **variable.** The variable itself may take on two or more **values.** Values can be defined as categories that constitute the variable.

For example, a ball (object) can be described in terms of its size (variable) as being "large" or "small" (values). In this instance, the variable SIZE is described rather crudely with the values large or small. SIZE also can be described utilizing more sophisticated values, such as centimeters to measure the diameter of the ball. Thus, the variable SIZE can describe the ball with either simple or exact categories. The different ways in which we choose to measure the size of the ball may reflect different **levels of measurement** and ultimately influence our choice of appropriate statistics. (See Chapter 5 for a more detailed discussion of levels of measurement and their role in statistical analysis.)

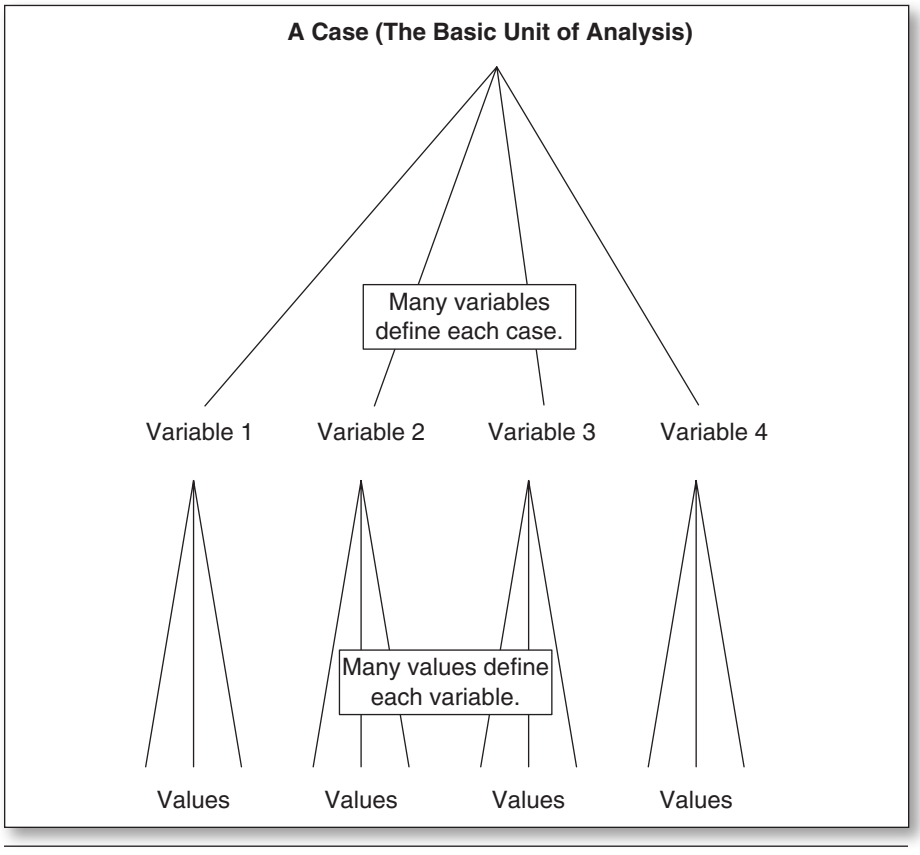
Because the ball has more than one characteristic, it can be described with more than one variable. For example, in addition to its size, a ball can have the following characteristics or variables: shape, color, construction material, and sport in which it is used (see Table 1.4). Thus, each case may be described using multiple variables, and each variable is likely to contain many values. This hierarchy is illustrated in Figure 1.1.

Table 1.4 How Variables Describe an Object

Name of Variable	*Possible Values*
Shape	Round or oval
Color	White or brown
Construction material	Plastic or leather
Sport	Golf or football

SOURCE: Newton and Rudestam (1999).

Figure 1.1 Illustration of Cases, Variables, and Values



SOURCE: Newton and Rudestam (1999).

Are There Rules I Should Use for Naming Variables?

Statistical software programs keep track of variables by using **variable names**. Typically, variable names are short mnemonics or numbers that refer to the variable. For example, the variable name for "age at first marriage" might be AGEWED. In this chapter we follow the rules set forth for naming variables in SPSS. Within SPSS, a variable name (a) must start with a letter and (b) may not contain any special characters such as blanks or commas. Thus, AGE WED would not be valid because it contains a space. Variables names such as 1VAR and VAR#1 would not be valid because the first starts with a number and the second contains a special character (#).On the other hand, VAR1 would be valid, as would V1 and VARIABLE1.

Some researchers prefer to identify variables strictly by numbers such as VAR001, VAR002, and so on. Others prefer to use mnemonics such as SEX, AGE, INCOME, and ETHNIC. Note that most statistical software programs also allow the user to attach a longer variable label to the variable name. Thus, the apparent confusion created by a long list of numbers (V1, V2, V3, etc.) can be alleviated by adding a complete list of variable labels. Our recommendation is to use what you find most comfortable; however, we have found that with standard scales, such as inventories assessing psychological or sociological constructs like depression, self-esteem, or alienation, it is better to use the number of the item in the scale as the variable name and provide more detailed information in the variable label. This helps you keep track of the items in the scale.

What Is Data Coding?

Note that in the foregoing example of the variables used to describe a ball, the majority of the values of the variables would be words like *plastic* and *rubber* rather than numbers. Measurement, however, was defined as the process of assigning numbers to characteristics of objects we observe. In preparing data for the computer, we assign numbers to the categories of variables that do not normally take numerical values. We call the process of assigning numbers to each variable's values the **coding process.** To illustrate, throughout this chapter we use questions from a hypothetical questionnaire that anyone might encounter.

Most surveys contain a "demographics" section that requests information about the respondents' age, sex, race, education, etc. Here is an example:

1. In the boxes below, write in your age at your last birthday.

The response to the question regarding respondent's age will be a number, not a word. Thus, the responses form the categories, with the youngest respondent's age forming the lowest value and the oldest respondent's age forming the highest value. A variable, such as AGE, is known as **self-coding** because the responses themselves create the numerical codes. For example, if the respondent was 36 years old, he or she would write 36 on the questionnaire. This self-coding numerical response would then be entered directly into a computer file, or it might be entered into a column on the right of the survey page and subsequently entered into a computer file.

Another question from our hypothetical questionnaire asks about the respondent's sex.

2. What is your sex? (circle one)

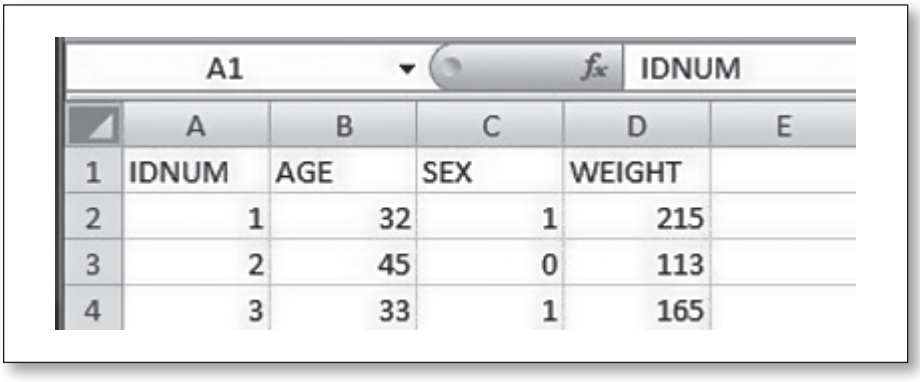
   Male 0

   Female 1

Note that in this question, the response (value) is a word (male or female), not a number as in question 1, which asked respondent's age. To code this variable, we have arbitrarily assigned 0 to male and 1 to female and included these codes to the right of each response. (We could just as easily have assigned "0" for female and "1" for male. The choice is up to the researcher within the limits of widely used conventions. (We are following a practice known as "dummy coding" of dichotomous variables, which will be discussed in Chapter 9.)

It is important to understand that the basic structure of the data we described earlier can be reproduced by a variety of software options, including spreadsheets, database managers, and, though not recommended, word processors. In the spreadsheet format, the number of digits in the variable would be largely irrelevant, because each variable would be entered into a cell in the spreadsheet. Table 1.5, a screenshot of a small Excel database, illustrates what this arrangement might look like in a spreadsheet program containing four variables that are coded into four columns, A, B, C, and D. Note that the

Table 1.5 Sample Excel Spreadsheet Illustrating Coding for Four Variables and Three Cases



A1 | IDNUM

	A	B	C	D	E
1	IDNUM	AGE	SEX	WEIGHT	
2	1	32	1	215	
3	2	45	0	113	
4	3	33	1	165	

rows (1, 2, 3, etc.) and columns (A, B, C, etc.) of a spreadsheet do not necessarily represent variable names or case identification numbers. We recommend including these within spreadsheet files.

What About Open-Ended Response Choices? How Are These Coded?

Sometimes a questionnaire item does not provide categories for all possible responses. For example, examine question 3 from the hypothetical questionnaire:

3. What race do you consider yourself? (circle one)

   White 1

   African American 2

   Other (specify) _____

When this question about race was constructed, it was assumed that the majority of the responses would fall into either the "White" or "African American" category. If the respondent's race is other than "White" or "African American," such as "Native American," he or she would write the response in the space provided:

Other (specify) _____

This type of response is known as **open-ended** or **nonstructured**. What if 5 respondents wrote in "Mexican American," 11 wrote in "Hispanic," and 7 wrote in "Native American"? Rather than just conclude that these were all "Other," you would probably want to identify these unique ethnic identities in your computer file. In order to do this, you would need to code the answers to an open-ended response category by completing the following steps.

Step 1. Record all the responses on a tally sheet to determine the names of races designated as "Other." It is also recommended that you record the number of times each race occurs (frequency). This might look like Table 1.6.

Table 1.6 Coding Frequency Data

Other: Specify	*Frequency Count*	N	*New Code*
Hispanic	11111 11111 1	11	3
Mexican American	11111	5	4
Native American	11111 11	7	5

For example, a respondent may have indicated that he or she is "Mexican American" or "Hispanic." If the total number of "Hispanics" is much greater than the remaining races listed in the "Other" category, you may decide to include "Hispanic" as a new code, giving it the "3" code as shown above.

Step 2. Assign additional codes to the new categories. For example, the tally sheet may resemble Table 1.6 in which Hispanic, Mexican American, and Native American have all been assigned code values.

Step 3. Finally, return to the questionnaire and write in the appropriate code for the "Other" category. For example, if a respondent had answered the question as "Native American" and you had coded "Native American" as "5," you would place a "5" in the right-hand column of the survey for later data entry.

3. What race do you consider yourself? (circle one)

White 1

African American 2

Other (specify) *Native American* Race_5_

How Do I Code Open-Ended Questions?

Open-ended questions are not as simple to code open-ended responses, illustrated previously. Consider questions 4 and 5 from a hypothetical questionnaire in which people are asked about their preference for burial or cremation as a method of body disposition and then asked the reason for their choice.

4. If you were to die within the next few years, do you feel positive, neutral, or negative about:

   A. Burial

      Positive 1
      Neutral 2
      Negative 3

   B. Cremation

      Positive 1
      Neutral 2
      Negative 3

5.

   A. Why burial? (within the next few years)
   B. Why cremation? (within the next few years)

The answers to question 5 may be longer and more complicated than the open-ended response regarding race. For example, to the question, "Why burial?" the following are several possible answers:

Burial gives me something to grieve over.

Burial is what has been practiced in my family.

My Catholic training in childhood taught me burial is the method of body disposition that most Catholics practice.

The idea of burning frightens me.

Because the answers are more complex, each response needs to be examined (i.e., content analyzed) to determine the underlying motivation. (A well-written introduction to content analysis is contained in Bruce L. Berg and Howard Lune's book *Qualitative Research Methods for the Social Sciences*, 2011.) In other words, the researcher asks, "Why did they say what they did?" These

motivations then become the categories into which the responses fall. Let's consider the four responses to the question, "Why burial?"

"Burial gives me something to grieve over." This answer might reflect the respondent's need for burial to aid in resolving grief. Therefore, "aid in grief resolution" may be an appropriate category for this response.

"Burial is what has been practiced in my family." It appears that burial has been a part of this person's family tradition. Therefore, "family tradition" may be an appropriate category for this response.

"My Catholic training in childhood taught me burial is the method of body disposition that most Catholics practice." This person's answer appears to suggest the influence of early religious training. Therefore, "religion" may be an appropriate category for this response.

"The idea of burning frightens me." It appears that this person's motive for choosing burial is a fear of cremation. Thus, "fear of burning" may be an appropriate category for the response.

As you examine each case to determine the underlying motivation, you may discover that some motivations are common to more than one respondent. For instance, if there are 100 cases in the study, there will not be 100 response categories for the question, "Why burial?" Instead, you may find that 10 people prefer burial for religious reasons, 20 people may be influenced by family tradition, 15 people fear burning, and so on. You may end up with several answers that are not at all alike and that do not seem to fit into any of the response categories created by the motivations. You may find it necessary to classify these types of responses as "miscellaneous motivations."

After you assign codes to the response categories that have been created, your tally sheet may resemble Table 1.7. Return to question 5 on each

Table 1.7 Coding Frequency Data

Motivations for Burial	*Frequency*	*Code*
Aid in grief resolution	11111 11111 1	1
Family tradition	11111 11111 11	2
Religion	11111 1111	3
Fear of burning	11111 11	4

questionnaire and record the category into which the response falls, as well as the code for that category.

What Are the Basic Guidelines of Data Organization?

There are numerous strategies for organizing data into a format that can be understood by a computer program that calculates statistics. Data can be entered in scientific notation, as currency or dates, and as strings of qualitative information such as names or open-ended responses. An individual with previous knowledge of this subject may have his or her own particular method for preparing and inputting data. For beginners, we suggest a strategy that has been proven, through experience, to maximize efficiency and minimize errors.

The following is a list of "basic guidelines of data organization." These guidelines are important. The longer you work with data entry, the more you will appreciate their value.

Guideline 1. It is generally best to use numeric codes for all your data, regardless of level of measurement.

In a questionnaire, that asks:

2. What is your sex? (circle one)

   Male 0

   Female 1

It may appear just as easy to code the values for sex as "M" and "F" as it is to code them "0" and "1." Such a practice, however, may prove troublesome when defining the data file for a statistical program. It is much easier to change variables that do not have numeric response choices, such as sex, to numerically coded variables before inputting the data than afterward. We strongly recommend the use of numeric codes as a standard procedure.

Guideline 2. All codes for a variable must be mutually exclusive.

This guideline has as much to do with good survey design as it does with data entry. By *mutually exclusive*, we mean that each case can be classified into one and only one category (value) on a particular variable. For example, consider the hypothetical questionnaire item below.

6. Are you currently married, not married, or do you have children?

   Currently married 1

   Not married 2

   Have children 3

It is quite probable that a respondent who has children will also fit into the "currently married" category. Because the respondent can fit into more than one category, the categories are not mutually exclusive. To satisfy the rule for mutually exclusive codes, it would be necessary to eliminate the category "have children" and add additional categories. The real issue here is that two variables are being assessed with one question (i.e., number of children and marital status). It is important to give your data collection strategy a thorough pretest to make sure you are aware of all the variables that can be derived from your instruments. A more traditional way to address the marital status questions is this:

6. Are you *currently* married, widowed, divorced, separated, or have you never been married? (circle one)

Currently married	1
Widowed	2
Divorced	3
Separated	4
Never married	5

Guideline 3. Each variable should be coded to obtain maximum information.

Let's consider the question regarding the respondent's marital status referred to in Guideline 2. When the values for respondent's marital status are "married" and "not married," respondents can only be classified into two broad categories. This provides only a minimum amount of information. All we know is that the respondent is either married or single. The five-category designation offers a more precise categorization, providing the researcher with more information for the marital status variable.

In another example of coding a variable to obtain maximum information, consider the question related to respondent's age:

1. In the boxes below, write in your age at your last birthday.

Let's suppose we change the wording of the above question to read:

1. How old were you on your last birthday?

Younger than 18	1
18–30	2
31–65	3
66 or over	4

When age is grouped into four categories, the exact age of the respondent is no longer available. Although the four-category classification may be helpful for identifying stages of adult development, it is not useful if the exact age of the subject is required. In addition, different theorists may have distinct views regarding the stages of adult development. A more technical consideration is the fact that by grouping age into categories we reduce the variability inherent in the ungrouped age distribution.

If the data are entered to represent the respondent's exact age, the computer can be given the task of creating an endless variety of age groupings. Once the data have been grouped and entered as above, the computer cannot determine the exact age of the respondent or restructure the data. Thus, it is possible to "collapse" data in numerous ways, but not possible to "expand" data that have been collapsed prior to data entry (a process called *recoding*). Note that one could also simply ask the respondent for his or her birthday and enter that information as a date variable. This would permit the calculation of the exact number of days the respondent had lived. Typically this amount of precision is not required; however, if the database was only composed on infants less than 1 year old, such information might be useful.

Guideline 4. For each case, there must be a numeric code for every variable.

Throughout the coding process, it is necessary for a code to be assigned to every variable. In some cases, there will be missing information, commonly referred to as **missing data** or **missing values**. For example, there may be no information regarding marital status for a particular case. Nevertheless, it is necessary to assign a code even when the datum is not available. Coding missing information will be explained in greater detail in the following section. At this point, suffice it to say that if missing information was not coded and the cell location in a spreadsheet was left blank, confusion could arise later in the analysis process. For example, if there was not a code for a particular variable, it might be difficult to determine whether or not a data entry error was made (whether the variable was inadvertently skipped during the data entry process) or if the respondent actually did not answer the question (a true missing value).

How Do I Define and Code Missing Data?

As we gather data for each case in our study, some items of information, for one reason or another, may not be available. This results in an incomplete set of data for some cases.

The following are some possible explanations for missing data:

The respondent refused to answer.

The respondent was not home.

During a phone interview, the respondent hung up.

The interviewer skipped three or four questions.

An experimental animal died halfway through the study.

A recording instrument failed.

Similarly, data collected by other persons—the US Census, for example—may not have certain pieces of information available for every case. Missing data must be included in the coding system. If the data are excluded, we have violated Guideline 4: "For each case, there must be a numeric code for every variable." In the SPSS statistical program, for example, any code can be designated as a missing value. A generally accepted convention is to use "9" for variables whose maximum value is no less than 1 or greater than 8, "99" for those variables whose values are greater than 8 but less than 98, and so on for three-digit (999) and larger variables.

For example, question 7 from the hypothetical questionnaire represents a one-column variable (there are fewer than nine response categories, excluding missing values):

7. How far did you go in school? (circle one)

Fewer than 8 years	1
Elementary school graduate	2
Some high school	3
High school graduate	4
Some college or AA degree	5
BA, BS, RN (college degree)	6
Some graduate school	7
Completion of advanced degree	8

Because there are fewer than nine response categories, we can designate "9" as the code for missing data. It is particularly important to point out here that assigning missing values during the data entry phase of a research project must

be followed by informing the statistical software of these missing value codes. The software won't automatically "know" that a "9" or "99" represents a missing value. During a *data definition session*, we inform the statistical software of our variable names, variable labels, value labels, and missing value codes.

How Do I Code Nonspecific Responses?

Sometimes a researcher believes it is necessary to analyze nonspecific responses instead of assigning them as missing data. For example, look at question 8 of the hypothetical questionnaire:

8. Do you believe there is a life after death? (circle one)

Yes	1
No	2
Undecided	8

"I don't know," "I'm not sure," and "Undecided" all represent nonspecific responses. Just as any code can be used to designate missing data, so too can any code be used to designate nonspecific responses. A generally accepted convention for nonspecific responses is to use "8" for one-digit variables, "98" for two-digit variables, and so on.

Exact coding decisions are up to the discretion of the researcher (within the specifications of the statistical program being used). A case in point is question 9 from the hypothetical questionnaire.

9. How often do you attend religious services? (circle one)

Never	0
Less than once a year	1
About once a year	2
Several times a year	3
About once a month	4
2–3 times a month	5
Nearly every week	6
Every week	7
Several times a week	8
No response	9

If the researcher wished to include a "don't know" code, creating 11 response categories (including the "don't know" response and the missing-value code 9), the variable would become a two-digit variable, and the codes would become the following:

Never	00
Less than once a year	01
About once a year	02
Several times a year	03
About once a month	04
2–3 times a month	05
Nearly every week	06
Every week	07
Several times a week	08
Don't know	98
No response	99

Certain coding conventions are not as universally followed as others, and the choice of codes is up to the researcher. However, following the guidelines suggested above, particularly those for the specification of missing values, will facilitate an understanding of your data structure choices among a wider range of researchers and reduce errors in the data analysis phase of your research.

How Do I Deal With Blanks and Zeros?

Blanks generally are not used to designate missing data. (Note that Guideline 4 of data organization would be violated; that is, "For each case, there must be a numeric code for every variable.") Zero codes also are not typically used to indicate missing data, because zeros frequently are used to represent "zero" of something. For example, in question 9, "How often do you attend religious services?" zero (0) represents "never."

We recommend that (a) blanks *never* be used in a data file and (b) zeros be used as missing value codes only when common sense dictates that a zero would be more appropriate than a 9.

This concludes our discussion of organizing data for computer input. In the following section, we present a detailed example using a new questionnaire, a codebook, and the basic guidelines of data organization.

A DETAILED EXAMPLE OF DATA ORGANIZATION AND CODING

When one enters data into a spreadsheet or directly into a statistical program, it may be difficult to retrieve the coding system without some directory giving the meaning of the various codes. As an example, consider the variable "Race/Ethnicity." Initially, a researcher may have decided to code Whites as "1," Hispanics as "2," Asians as "3," and African Americans as "4." Without labels linking each code with its respective category, however, it becomes impossible to know that Hispanics were originally coded as "2." Such information is contained in a document known as a **codebook**. A codebook for the sample questionnaire can be found at the end of this chapter.

In this section, we want to reinforce your understanding of the basic rules of data organization by using a questionnaire to construct a codebook. A codebook is an index to your coding system. In it, you provide the names of your variables; a description of each variable (a short reference or a detailed description if necessary); and the codes and code categories you have devised, including the codes for missing and nonspecific responses. In addition, depending on the form of the data (e.g., contained in an Excel spreadsheet), you might include the column location of each variable.

Codebooks vary widely in structure and content, but most major research organizations, such as the National Opinion Research Center, the Gallup and Roper organizations, the US Census, and the Inter-university Consortium for Political and Social Research distribute their data accompanied by codebooks. Without these, the structure of the data would remain a mystery. Sometimes codebooks include the complete text of every question and a univariate data distribution; in other cases, they are simpler. The following pages present a five-item questionnaire. Alongside each page of the questionnaire is a blank codebook. Try to construct the codebook based on the information presented in the questionnaire. For assistance in coding, return to the section "What Are the Basic Guidelines of Data Organization?" where data organization for computer input is discussed. Following is a list of helpful guidelines:

1. *Variable name*. We will adopt the specifications of SPSS and create variable names with no blank spaces between characters, using only numbers and letters, and always starting with a letter. For example,
   a. The variable name for respondent's ID number can be "ID" or "IDNUM" but not "ID NUM."
   b. Respondent's age can be "V01" or "AGE" or "RESPONDENTS_AGE" but not "RESPONDENT'S AGE." (Note that the underline character ("_") is acceptable in SPSS.)

2. *Variable description (label)*. This is a brief description of the variable. It should enable anyone to readily identify the questionnaire item. For example, "attitudes toward pornography laws" is an adequate variable description to identify questionnaire item 1.
3. *Value codes and value labels*. These are the names of the response categories and their respective codes. You will note that all response categories for the remaining questions have been coded, with the exception of respondent's ID number and question 2, which are both self-coding. Typically self-coded variables are not provided value labels.

The Questionnaire

1. Which of these statements comes closest to your feelings about pornography laws? (circle one)

   There should be laws against the distribution of pornography whatever the age 1

   There should be laws against the distribution of pornography to persons under 18 2

   There should be no laws forbidding the distribution of pornography 3

2. On the average day, about how many hours do you personally watch television? Enter number of hours: _____

3. Here are four statements regarding the role of the working mother. Please circle whether you *strongly agree, agree, disagree*, *strongly disagree*, or *don't know* with each statement.

   A. A working mother can establish just as warm and secure a relationship with her children as a mother who does not work.

Strongly agree	1
Agree	2
Disagree	3
Strongly disagree	4

   B. It is more important for a wife to help her husband's career than to have one herself.

Strongly agree	1
Agree	2

Disagree 3

Strongly disagree 4

C. A preschool child is likely to suffer if his or her mother works.

Strongly agree 1

Agree 2

Disagree 3

Strongly disagree 4

D. It is much better for everyone involved if the man is the achiever outside the home and the woman takes care of the home and family.

Strongly agree 1

Agree 2

Disagree 3

Strongly disagree 4

4. In 2008, Obama ran for president on the Democratic ticket against McCain for the Republicans. Do you remember for sure whether or not you voted in that election?

Voted 1

Eligible but did not vote 2

Ineligible 3

A. IF VOTED: Did you vote for Obama, McCain, or for another candidate?

Obama 1

McCain 2

Other candidate (Specify: _____) _____

B. IF DID NOT VOTE OR WAS INELIGIBLE:

Whom would you have voted for, for president, if you had voted?

Obama 1

McCain 2

Other candidate (Specify: _____) _____

5. For each area of life mentioned below, circle the number that shows how much satisfaction you get from that area.

   A. The city or place you live in

A very great deal	1
A great deal	2
Quite a bit	3
A fair amount	4
Some	5
A little	6
None	7

   B. Your nonworking activities—hobbies and so on

A very great deal	1
A great deal	2
Quite a bit	3
A fair amount	4
Some	5
A little	6
None	7

   C. Your family life

A very great deal	1
A great deal	2
Quite a bit	3
A fair amount	4
Some	5
A little	6
None	7

   D. Your friendships

A very great deal	1
A great deal	2

Quite a bit	3
A fair amount	4
Some	5
A little	6
None	7

E. Your health and physical condition

A very great deal	1
A great deal	2
Quite a bit	3
A fair amount	4
Some	5
A little	6
None	7

The Codebook

Question Number	*Variable Name*	*Variable Description (Label)*	*Value Codes & Value Label*
Identification Number	IDNUM	Respondent's Identification Number	Self-coding 999 Missing value
1	V01[1]	Attitudes toward pornography laws	1 Porno laws for all ages[2] 2 Porno laws for under 18 3 No porno laws 8 Don't know 9 Missing value[3]
2	V02	Hours spent watching TV	Self-coding 98 Don't know 99 Missing value

(Continued)

(Continued)

Question Number	*Variable Name*	*Variable Description (Label)*	*Value Codes & Value Label*
3A	V03	Working mother can have warm relationship with children	1 Strongly agree 2 Agree 3 Disagree 4 Strongly disagree 8 Don't know 9 Missing value
3B	V04	Wife helps husband's career rather than her own	1 Strongly agree 2 Agree 3 Disagree 4 Strongly disagree 8 Don't know 9 Missing value
3C	V05	Preschool child suffers if mother works	1 Strongly agree 2 Agree 3 Disagree 4 Strongly disagree 8 Don't know 9 Missing value
3D	V06	Man works—woman cares for home and family	1 Strongly agree 2 Agree 3 Disagree 4 Strongly disagree 8 Don't know 9 Missing value

Question Number	*Variable Name*	*Variable Description (Label)*	*Value Codes & Value Label*
4	V07	2008 presidential election	1 Voted 2 Eligible but did not vote 3 Ineligible 8 Don't know 9 Missing value
4A	V08	If voted	1 Obama 2 McCain 3 Other (specify) 8 Don't know 9 Missing value
4B	V09	If did not vote or ineligible, for whom would you have voted	1 Obama 2 McCain 3 Other (specify) 8 Don't know 9 Missing value
5A	V10	Satisfaction from place you live	1 A very great deal 2 A great deal 3 Quite a bit 4 A fair amount 5 Some 6 A little 7 None 8 Don't know 9 Missing value

(Continued)

(Continued)

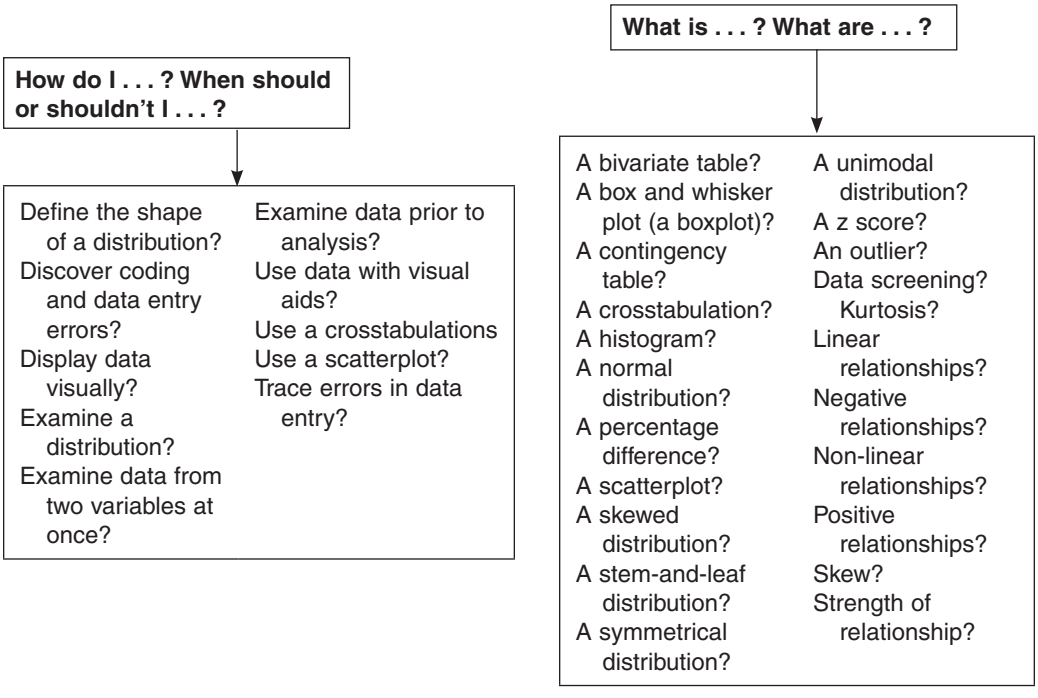
Question Number	*Variable Name*	*Variable Description (Label)*	*Value Codes & Value Label*
5B	V11	Satisfaction from hobbies	1 A very great deal 2 A great deal 3 Quite a bit 4 A fair amount 5 Some 6 A little 7 None 8 Don't know 9 Missing value
5C	V12	Satisfaction from family life	1 A very great deal 2 A great deal 3 Quite a bit 4 A fair amount 5 Some 6 A little 7 None 8 Don't know 9 Missing value
5D	V13	Satisfaction from friendships	1 A very great deal 2 A great deal 3 Quite a bit 4 A fair amount 5 Some 6 A little

Question Number	*Variable Name*	*Variable Description (Label)*	*Value Codes & Value Label*
			7 None 8 Don't know 9 Missing value
5E	V14	Satisfaction from health	1 A very great deal 2 A great deal 3 Quite a bit 4 A fair amount 5 Some 6 A little 7 None 8 Don't know 9 Missing value[4]

NOTES:

1. We chose to use alphanumeric variable names in this codebook (V01). Variable names that suggest the name of the variable itself may also be used. For example, we used IDNUM as the variable name for respondent's ID number, and V02 could have been named TVHOURS.
2. The value labels are a description of the response categories and their respective codes. Some statistical packages may limit the length of labels.
3. We reserve "8" to designate nonspecific responses and "9" for missing data. See Guideline 4 of data organization.
4. The codes are all numeric and thus satisfy Guideline 1 of data organization.

How Do I Examine Data Prior to Analysis?



Level: Beginning

Focus: Instructional

CHAPTER 2

How Do I Examine Data Prior to Analysis?

Imagine mounting a large data collection effort, entering all the data into a computer following the guidelines in the first chapter, and then facing the question, "What do I do now?" Some researchers might be tempted to jump immediately to the testing of hypotheses using multivariate techniques, whereas others might not know where to start. This chapter addresses the question of where to start with data analysis. There are two goals of this chapter. First, the procedures we suggest, though relatively easy to follow and understand, serve as the second step in a quality control process that protects data analysts from disaster. (The first step, which occurs at the data entry point, was discussed in Chapter 1.) Second, the blueprint for preliminary data inspection that we outline here is central to the plan of this entire book. The well-prepared data analyst should be familiar with the basics outlined here before proceeding further.

WHEN SHOULD I SCREEN DATA?

It is naive to believe that once the data are collected, all the hard work is done because the computer will simply "crunch the numbers" and answer all our questions. In fact, just as with qualitative analysis, one needs to get a feel for one's data and become comfortable with them before jumping into full-scale hypothesis testing. This process starts from the ground up, first with the examination of one variable at a time (univariate analysis) and then building toward more complex statistical applications (multivariate analysis), if necessary and appropriate.

Examination of the characteristics of data is partly a visual activity, and we emphasize the visual inspection of data in this and other sections. A thorough

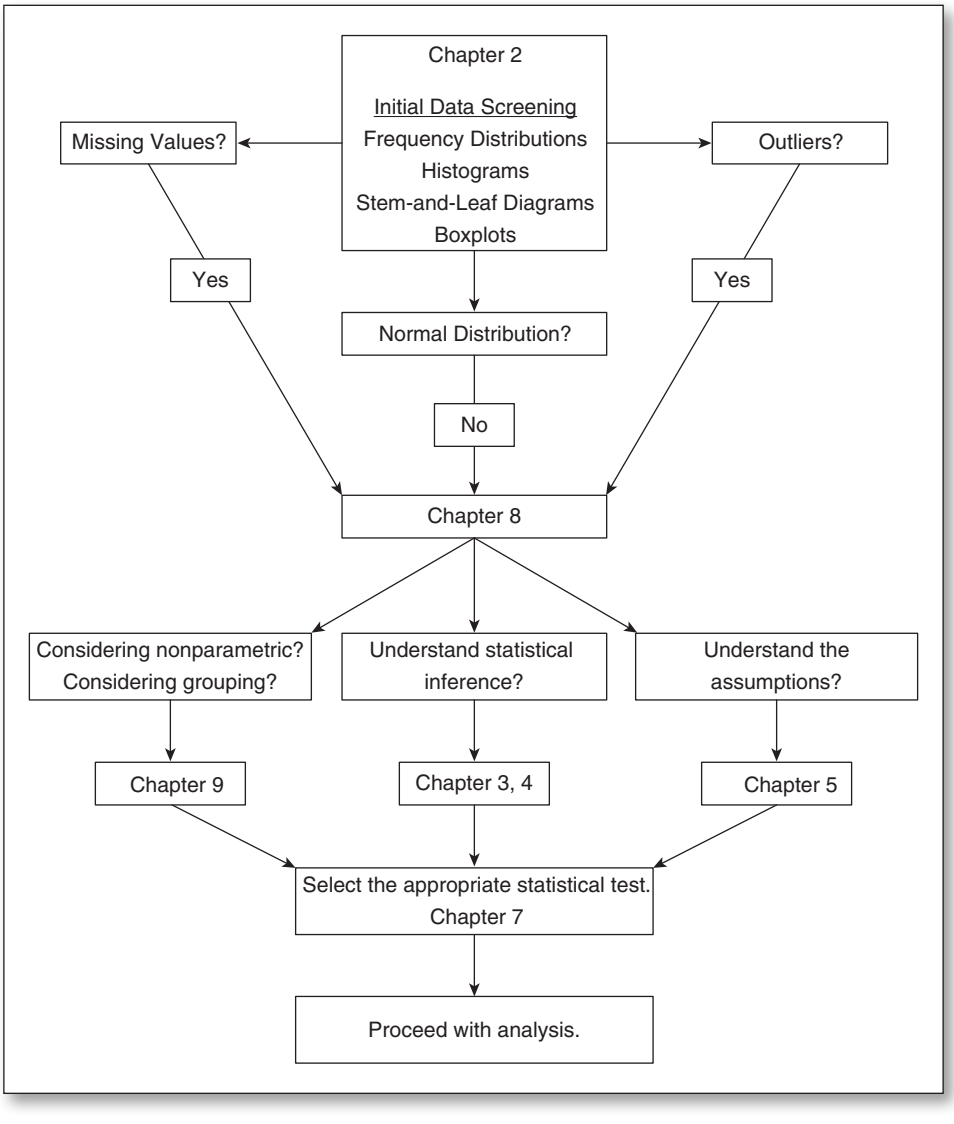
inspection of data also makes use of tabular and statistical manipulations. When we thoroughly understand the properties of our data, many of the questions regarding the ability of the data to meet the assumptions of sophisticated statistical analyses will be answered. In this chapter, we begin by suggesting strategies for examining the distribution of a single variable, particularly the ability of that variable to approximate a normal distribution. We follow this by suggesting methods for examining bivariate distributions. This process is likely to indicate issues or problems concerning the data that require some "adjustments" before moving on. Such adjustments include methods for detecting and correcting for missing observations and extreme scores or outliers. These topics are discussed in greater detail in Part III: Issues Related to Variables and Their Distributions. The flowchart in Figure 2.1 describes the data-screening process and the chapters in which specific issues are considered.

The initial step in the exploration of a data set is to examine the distribution of every variable that composes that data set. The starting point of this process is most typically the examination of the frequency distributions of the variables in question. This is necessary regardless of whether the variables are categorical or continuous and regardless of what statistical operations we ultimately conduct.

We present frequency distributions throughout this book, and most computer programs provide distributions with a similar structure that can be modified by the user. We will be using distributions as produced by hand and by IBM-SPSS Statistics Version 19®. For example, in Table 2.1 we present the distribution of "Age When First Married [AGEWED]" from the National Opinion Research Center General Social Survey of 2006, as produced by IBM-SPSS®. (Printouts from IBM-SPSS may differ slightly in their appearance from the reproductions in this book, but the data contained will be identical.) Note that the table has a column showing the valid codes (the ages) and the missing values (NAP for "Not Applicable"; NA for "No Answer"). The 3,342 NAP codes stand for the "not applicable" cases, including those who were never married and those, due to the sampling design of the General Social Survey, who were not scheduled to receive this question. The table contains columns of frequencies, percentages, valid percentages, and cumulative percentages. The difference between the percentages and valid percentages results from the treatment of missing values; the percentage column includes them, but the valid percentage column does not.

There a number of things to look for during a first glance at a frequency distribution such as that in Table 2.1. First, we should look for codes we do not understand or that we may not wish to include among the valid cases.

Figure 2.1 The Data-Screening Process



SOURCE: Newton and Rudestam (1999).

This is particularly true when conducting secondary analysis of an existing database, as represented here by our use of the General Social Survey.

Table 2.1 Distribution of Age When First Married

Age When First Married		*Frequency*	*Percent*	*Valid Percent*	*Cumulative Percent*
Valid	13	2	0.0	0.2	0.2
	14	3	0.1	0.3	0.4
	15	4	0.1	0.3	0.8
	16	25	0.6	2.2	2.9
	17	46	1.0	4.0	6.9
	18	117	2.6	10.1	17.0
	19	98	2.2	8.4	25.4
	20	111	2.5	9.6	34.9
	21	139	3.1	12.0	46.9
	22	95	2.1	8.2	55.1
	23	83	1.8	7.1	62.2
	24	72	1.6	6.2	68.4
	25	71	1.6	6.1	74.5
	26	54	1.2	4.6	79.2
	27	53	1.2	4.6	83.7
	28	32	0.7	2.8	86.5
	29	23	0.5	2.0	88.5
	30	25	0.6	2.2	90.6
	31	15	0.3	1.3	91.9
	32	15	0.3	1.3	93.2
	33	12	0.3	1.0	94.2
	34	11	0.2	0.9	95.2

Age When First Married		*Frequency*	*Percent*	*Valid Percent*	*Cumulative Percent*
	35	11	0.2	0.9	96.1
	36	3	0.1	0.3	96.4
	37	5	0.1	0.4	96.8
	38	5	0.1	0.4	97.2
	39	4	0.1	0.3	97.6
	40	9	0.2	0.8	98.4
	42	2	0.0	0.2	98.5
	43	2	0.0	0.2	98.7
	44	3	0.1	0.3	99.0
	45	1	0.0	0.1	99.1
	46	2	0.0	0.2	99.2
	48	2	0.0	0.2	99.4
	52	2	0.0	0.2	99.6
	56	1	0.0	0.1	99.7
	65	1	0.0	0.1	99.7
	70	1	0.0	0.1	99.8
	90	1	0.0	0.1	99.9
	DK	1	0.0	0.1	100.0
	Total	1,162	25.8	100.0	
Missing	NAP	3,342	74.1		
	NA	6	0.1		
	Total	3,348	74.2		
Total		4,510	100.0		

(The complete data from the General Social Survey, all 55,087 cases from 1972 through 2010, is available at **http://sda.berkeley.edu/** and may be either downloaded or analyzed online. This site also archives data from numerous other studies.) When examining the distribution of age at first marriage, note the DK (Representing "Don't Know") code that has not been assigned as a missing value. While this represents only a single case, we would want to remove this case from the list of valid cases by reassigning the DK code value (98) as a missing value.

Second, we should attain a preliminary sense of the shape of the distribution. It appears that this distribution has a few cases that were married very young (16 or younger, 2.9%) and a few cases married after 40, up to an age of 90 (1.6%). This long trailing of a few cases between 40 and 90 creates what is referred to as "positive skew," because the upper tail of the distribution tends to be pulled outward by a few extreme cases. The term **skew,** as used to describe a distribution, generally refers to the distribution's deviation from a perfectly symmetrical shape. It is easy to see from this distribution that most people were first married between the ages of 17 and 25, but a few were first married in their 40s or 50s and even older; hence the positive skew.

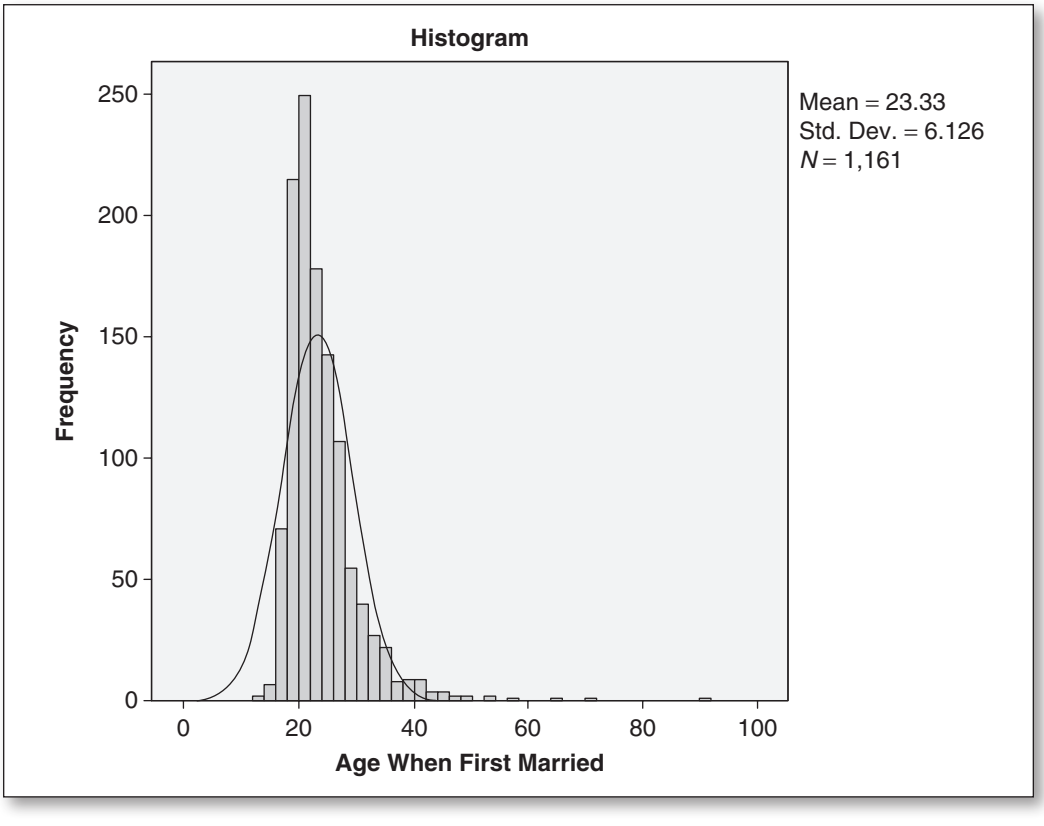
The third issue that researchers should be aware of when first examining their data is the possibility that errors in coding and recording were made. This is part of what researchers call "data cleaning." Sometimes data just don't make sense. For example, the distribution of age at first marriage shows that the youngest person married at the age of 13. This is probably a valid case, but if the youngest age was reported as 5 years, we might become suspicious of a data entry error. We might also question whether or not the person who was first married at 90 is a valid response; however, with secondary analysis we may never know the answer to that question and will have to accept the data at face value.

A good policy is to routinely produce the frequency distribution of every variable, even the identification numbers of cases, looking for these obvious errors. (When an ID number occurs twice, it may indicate that some data have been inadvertently entered twice or a data entry error was made, a very common occurrence with larger data sets.) Not all data entry errors are as easy to detect as obviously suspect codes or repeated ID numbers. If a 19 instead of a 29 was entered for marriage age, it is almost a certainty that the error will never be detected. The best precautions are first to enter the data carefully and second to return to the original data and trace obvious errors from there.

HOW DO I DISPLAY MY DATA VISUALLY?

An examination of the frequency distribution might be sufficient for some types of research; however, if your goal is to utilize the variable in procedures involving statistical inference, you probably will also want to examine visual representations of the data. Most statistical packages create graphs, and some provide options, such as superimposing a normal curve over the histogram of the distribution, to assist with the evaluation of the shape of the distribution, as is shown in Figure 2.2 for age at first marriage. Note that we have removed the one "Don't Know (DK)" response from this distribution as it erroneously inflates positive skew.

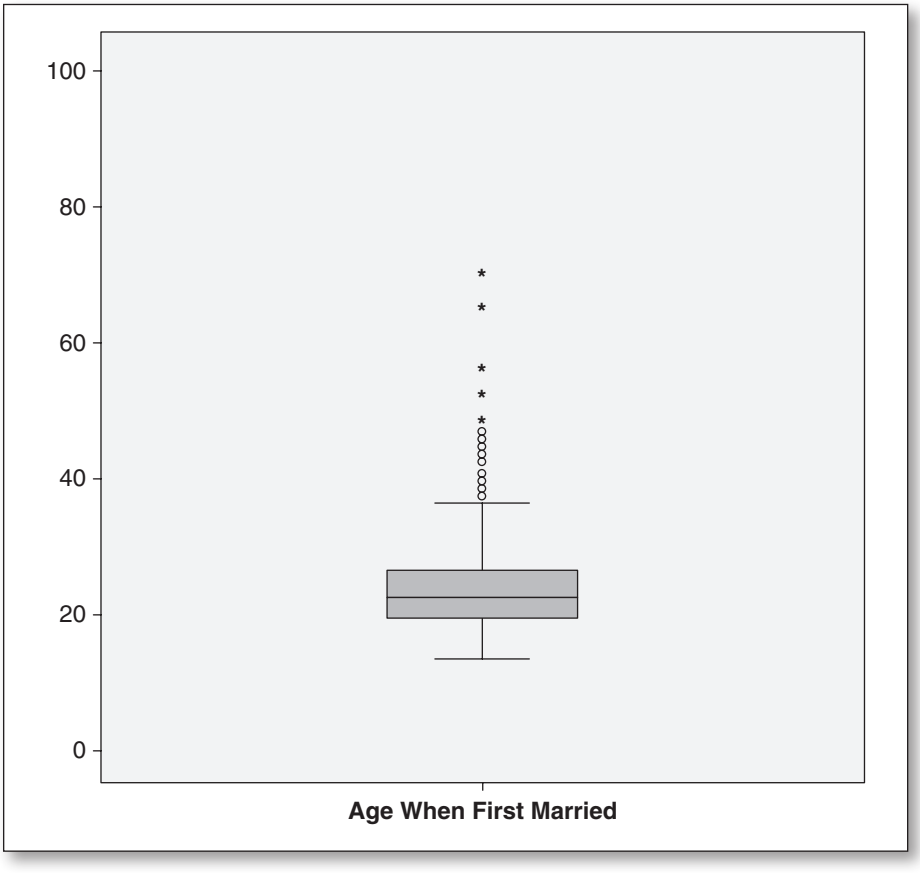
Figure 2.2 Histogram of Age When First Married With Superimposed Normal Curve



The histogram clearly shows that the right tail of the distribution exceeds that of the superimposed normal curve (positive skew) and that the distribution is more "peaked" than a normal distribution. The general "peakedness" of a distribution is called **kurtosis.** Very flat distributions are called **platykurtic**; very peaked distributions, such as this, are called **leptokurtic**; and distributions approximating a bell-shaped normal curve are called **mesokurtic**. Thus, the histogram helps us to see that the distribution of age of first marriage is both positively skewed and leptokurtic.

A number of other choices exist for presentation of a univariate distribution of continuously distributed data. Two of these are the boxplot, or box and whisker plot, and the stem plot, or stem-and-leaf diagram. We illustrate each in the following, beginning with the boxplot for the age of first marriage variable in Figure 2.3.

Figure 2.3 Boxplot for Age When First Married



What Is a Box and Whisker Plot?

A **box and whisker plot** contains a box whose boundaries represent cutoff points for the upper and lower quartiles of the distribution. This means that the middle 50% of the cases is contained within the box and the upper and lower 25% (below the first and above the third quartiles) are excluded. The horizontal line within the box represents the median, or the value of the case in the exact middle of the distribution (the point at the 50th percentile). The height of the box then can be interpreted as the distance between the first and third quartiles. The lines extending from the top and bottom of the box are called **whiskers** and represent the largest and smallest values that are not considered outliers (i.e., extreme scores).

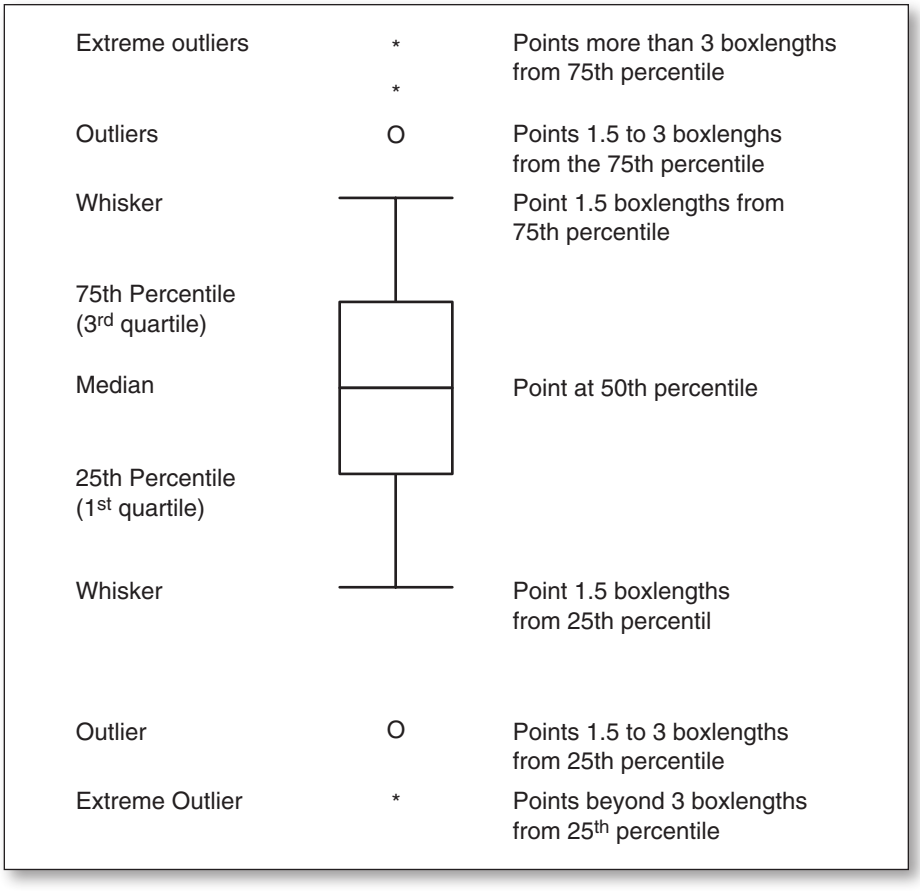
Outliers, represented with circles (O) and asterisks (*), are of two types: outliers (the circles) and extreme outliers (the asterisks). Outliers are values that range from 1.5 to 3 box lengths above or below the edges of the box; values more than three box lengths from the edge are called extreme outliers. Different authors may define outliers differently, so be aware that this is only one recommended definition. Some computer programs (including IBM-SPSS) also provide the identification numbers of the cases that are considered outliers, but we have excluded those here. The inclusion of case numbers help the researcher return to the original data to examine the case, if necessary. A more thorough discussion of how to deal with outliers can be found in Chapter 8. A diagram of this general framework for presenting box and whisker diagrams is presented in Figure 2.4.

Note that the boxplot for age at marriage (Figure 2.3) shows the median at approximately 22, with the upper (75th) and lower (25th) percentiles at 26 and 19, respectively. There is also a large number of outliers (41, to be exact). These are cases beyond a value of 36.5, which would be 1.5 box lengths above the box boundary of 25. We obtained the value of 36.5 by first calculating a box length of 7 (26 – 19 = 7). Multiplying this value by 1.5 ($7 \times 1.5 = 10.5$) and adding this to the value of the upper box boundary (26), we get $26 + 10.5 = 36.5$. Values beyond 36.5 are thus classified as outliers, and values beyond 43, or 3 box lengths ($7 \times 3 = 21$; $25 + 21 = 46$), are extreme outliers. Thus, the boxplot helps us visualize the distribution and clearly identifies the reason for the positive skew of this distribution, that is, the large number of persons married beyond the age of 36, including our one lucky 90-year-old.

What Is a Stem-and-Leaf Diagram?

A **stem-and-leaf diagram** is another way to visualize the distribution of a variable. Stem-and-leaf diagrams replace the bars of a histogram with values

Figure 2.4 The Box and Whisker Plot (boxplot)



SOURCE: Newton and Rudestam (1999).

obtained from the data. A stem-and-leaf diagram is best explained with an example. We present the stem-and-leaf diagram of the age at marriage variable in Figure 2.5.

Note that each stem represents one of the age categories, from 13 to 36, and that each leaf, represented by a 0, represents 2 cases. The 41 outliers (values ≥37) are not included. Among other things, one can see that without the outliers, the distribution appears more symmetrical but still maintains its leptokurtic character. In Figure 2.6, we present a number of small graphs that depict the language that statisticians use to describe the shape and form of a distribution.

Figure 2.5 Stem-and-Leaf Diagram of Age When First Married

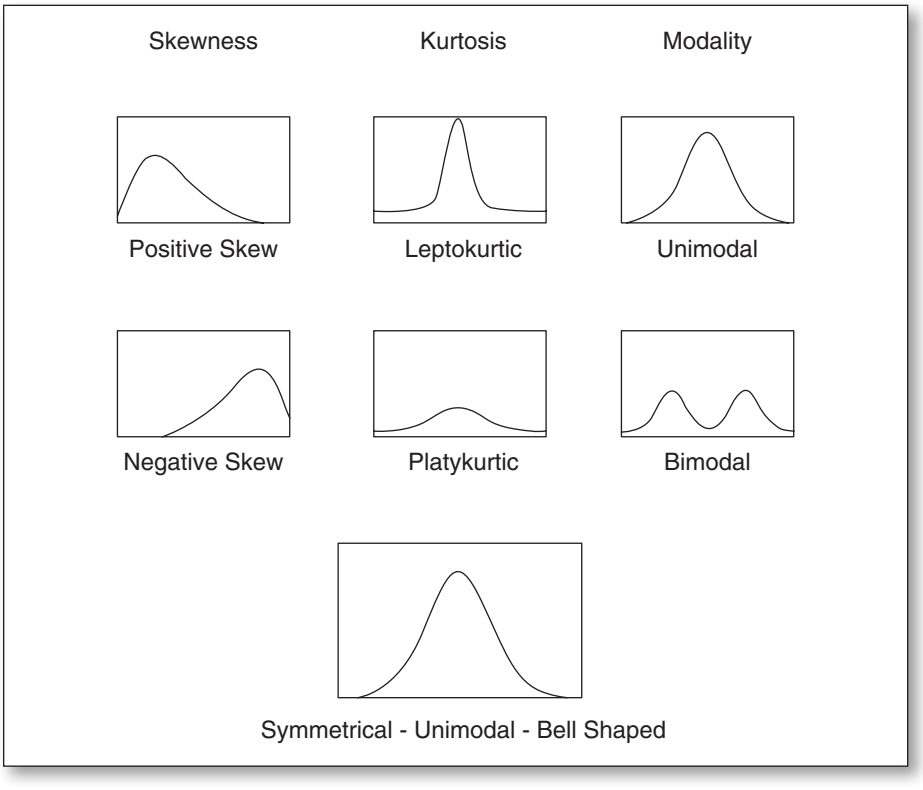
```
Frequency   Stem &  Leaf
  2.00      13   .  0
  3.00      14   .  0
  4.00      15   .  00
 25.00      16   .  0000000000
 46.00      17   .  00000000000000000000
117.00      18   .  000000000000000000000000000000000000000000000000000000000
 98.00      19   .  00000000000000000000000000000000000000000000000000
111.00      20   .  000000000000000000000000000000000000000000000000000000
139.00      21   .  0000000000000000000000000000000000000000000000000000000000000000000000000
 95.00      22   .  00000000000000000000000000000000000000000000000
 83.00      23   .  000000000000000000000000000000000000000000
 72.00      24   .  0000000000000000000000000000000000
 71.00      25   .  0000000000000000000000000000000000
 54.00      26   .  000000000000000000000000
 53.00      27   .  000000000000000000000000
 32.00      28   .  00000000000000000
 23.00      29   .  00000000000
 25.00      30   .  0000000000000
 15.00      31   .  0000000
 15.00      32   .  0000000
 12.00      33   .  000000
 11.00      34   .  00000
 11.00      35   .  00000
  3.00      36   .  0
 41.00 Extremes (>=37)

Stem width: 1
Each leaf: 2 case(s)
```

WHAT IS A NORMAL DISTRIBUTION?

The term *normally distributed* appeared a number of times in the preceding discussion and will appear frequently throughout this book. A **normal distribution** is a theoretical probability distribution and is a special case of a symmetrical,

Figure 2.6 Distribution Shapes and Forms

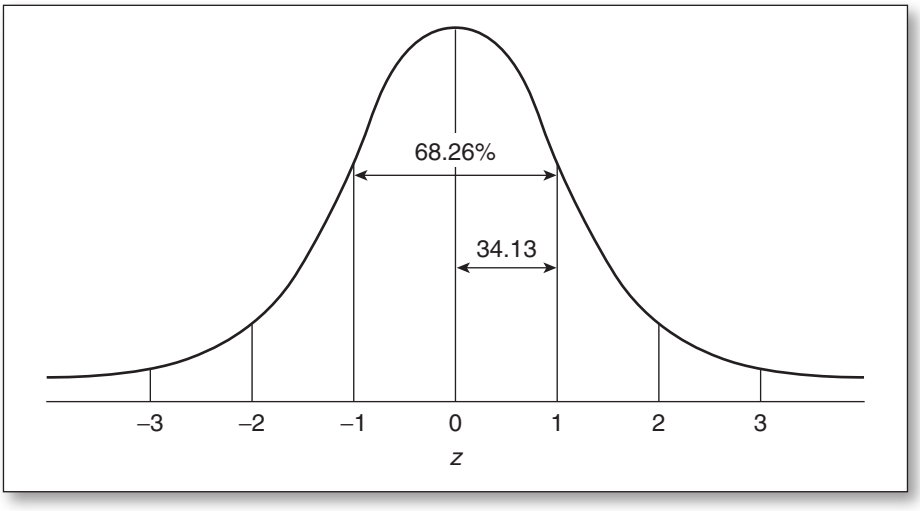


SOURCE: Newton and Rudestam (1999).

unimodal, bell-shaped curve. This curve may represent the distribution of real-world phenomena, but as shown in the examples above, many distributions deviate markedly from a symmetrical, bell-shaped structure.

Virtually all statistics texts provide examples of normal curves and contain a table that defines the area underneath a normal curve. These areas are located through the use of *z* scores and directly correspond to probabilities. A *z* score is a "transformation" of a distribution in such a way that the mean of the distribution becomes zero and the standard deviation becomes 1. When applied to a normal distribution, the *z* score transformation standardizes the distribution, and the term *standard normal distribution* is used to describe these curves. (A *z* transformation may be applied to any distribution, but it does not create a normal distribution from an original distribution that is not normal.) Thus, a *z* score of +1 indicates the point on the horizontal axis that is 1 standard

Figure 2.7 A Standard Normal Distribution



SOURCE: Newton and Rudestam (1999).

deviation above the mean, and a *z* score of –2 indicates a point 2 standard deviations below the mean. Figure 2.7 presents an example of a normal curve represented as a standard normal distribution.

Note that the horizontal axis in Figure 2.7, labeled *z*, has positive and negative values around a mean of zero. These *z* scores represent standard deviations above and below the mean. In a normal distribution, the area between the mean and the *z* scores is known and is defined by the normal curve table. For example, in Figure 2.7, the area between the mean and a *z* score of +1 is .3413 (34.13%). Because the curve is perfectly symmetrical, the area between the mean and a *z* score of –1 is also .3413. The area between a *z* score of +1 and –1 is thus the sum of these two areas, or .6826. Because the normal distribution is a probability distribution, the areas between any two points on the curve can be interpreted as probabilities. This point will become extremely important in our discussion of the logic of statistical inference in Part II of this book.

HOW DO I PREPARE TO ANALYZE CATEGORICAL DATA?

Our previous discussion refers entirely to data that are represented on a continuous numeric scale, such as age. What about data that are categorical, such as marital status or political preference?

First, many of the same considerations discussed in reference to a continuous variable, such as age or age of first marriage, also apply to variables that are categorical. The distributions of these variables should be carefully examined for coding errors and the assignment of missing values. While the concepts of skew and normality do not generally apply for categorical variables, it is important to assess the distribution of cases into the various categories. For example, Table 2.2 presents the distribution of a variable from the 2006 General Social Survey that assesses the respondents' attitudes toward extramarital sex (XMARSEX).

Note two important considerations about this variable. First, as seen with the age of first marriage variable, the DK (Don't Know) responses have been included and should probably be assigned as missing. Second, note that almost everyone in this sample of Americans (79.3%) believes that extramarital sex is always wrong, and only 1.8% ascribe to the "not wrong at all" position. Thus, the very small frequency of cases in this category needs to be taken into account when conducting analyses. Possible options include combining these with another category, such as the "sometimes wrong" group of 121, or eliminating

Table 2.2 Distribution of Attitudes Toward Extramarital Sex

SEX WITH PERSON OTHER THAN SPOUSE					
		Frequency	*Percentage*	*Valid Percentage*	*Cumulative Percentage*
Valid	ALWAYS WRONG	1,582	35.1	79.3	79.3
	ALMST ALWAYS WRG	236	5.2	11.8	91.1
	SOMETIMES WRONG	121	2.7	6.1	97.2
	NOT WRONG AT ALL	36	0.8	1.8	99.0
	DK	20	0.4	1.0	100.0
	Total	1,995	44.2	100.0	
Missing	NAP	2,507	55.6		
	NA	8	0.2		
	Total	2,515	55.8		
Total	4,510	100.0			

them entirely (probably not recommended in this example, but perhaps a valid option in others). The main point is that understanding the characteristics of every distribution in a database, your own or one acquired through secondary sources, is a critical first step in successful data analysis.

HOW DO I EXAMINE TWO VARIABLES AT THE SAME TIME?

Once we have a reasonable picture of the univariate distributions of our variables, we can begin to consider how these variables may work together. In other words, we can begin to think about relationships between variables. There are two basic, preliminary strategies for doing this, which depend on the characteristics of the data. For two *categorical* variables with a small number of categories, either discrete or orderable discrete, we recommend a **crosstabulation**. For two *continuous* variables, the **scatterplot** is the method of choice. When one variable is continuous and one discrete, we may choose to construct side-by-side boxplots or back-to-back stem-and-leaf plots, or we might use clustered bar charts. We briefly describe each of these strategies next.

How Do I Use a Crosstabulation to Examine Categorical Variables?

A **crosstabulation** (also called a bivariate table or contingency table) contains the joint distribution of two categorical variables. The categories of each variable are laid out in the form of a square or rectangle containing rows (representing the categories of one variable) and columns (representing the categories of the second variable). Thus, if we wished to examine the bivariate relationship between sex and attitudes toward abortion, our crosstabulation might appear as in Table 2.3.

The entries in each cell will be the frequencies representing the number of times that each pair of values (one from each variable) occurs in the sample. These are called **cell frequencies**. The entries that occur at the end of each row or column are simply sums across the rows or columns and are called **marginal frequencies**. These represent the total number of times the category of the row variable or column variable occurred. By adding either the row marginal frequencies or the column marginal frequencies, we obtain the total N (total sample size). By widespread agreement, the conventional way to construct a table is with the independent variable at the top and the dependent variable at the side of the table. Thus, if we believe that attitudes toward abortion are

Table 2.3 Example of Crosstabulation Between Sex and Attitudes Toward Abortion

Attitude Toward Abortion	*Sex*	
	Male	*Female*
Always wrong		
Almost always wrong		
Never wrong		

SOURCE: Newton and Rudestam (1999).

partly a function of sex (the only logical choice in this instance), then sex is placed at the top of the table and becomes the column variable.

Our goal with crosstabulation is, first, to get a sense of how the cases distribute themselves into the cells of the table (i.e., their joint distribution) and, second, to get a sense of whether or not the two variables are related. For example, examine the three samples shown in Table 2.4, each giving a distribution of cases in the sex-by-abortion crosstabulation.

The first sample in Table 2.4 illustrates a situation in which the examination of the relationship between the two variables would probably be unreasonable because, for whatever reason, the sample was composed mostly of women. With only 10 men in the sample, it probably would be unreasonable to conclude that sex and attitudes toward abortion are related or to calculate measures of statistical association. The second sample shows a situation in which the cases do distribute among the cells and no cell frequency is exceedingly

Table 2.4 Sample Outcomes for Attitudes Toward Abortion, by Sex

Attitude Toward Abortion	*Sample 1*		*Sample 2*		*Sample 3*	
	M	*F*	*M*	*F*	*M*	*F*
Always wrong	4	804	380	760	380	380
Almost always wrong	3	477	70	140	310	80
Never wrong	3	209	50	100	310	40
Total	10	1,490	500	1,000	1,000	500

SOURCE: Newton and Rudestam (1999).

small, but the variables are unrelated. The third sample shows a situation in which the cases distribute among the cells and the variables are related.

How could we determine that the variables are unrelated in sample 2 but related in sample 3, when in fact one might conclude exactly the opposite (in the second sample, twice as many females as males said abortion was always wrong, but in the third table the numbers are equal)? The answer lies in how we standardized the samples by calculating percentages. When the total numbers of frequencies in the categories of the independent variable are not identical, little information regarding the relationships can be gained by comparing these frequencies. Comparing percentages, on the other hand, may yield valuable information.

We may calculate percentages in three ways: within columns, within rows, or with the total *N*. In fact, only one of these ways provides meaningful answers to the question of the relationship between the two variables. This is to calculate the percentages within the columns. In other words, we want to standardize within the categories of the independent variable. When working within columns, the column total becomes the base number used to calculate the column percentage. For example, to find the percentage of males among those believing that abortion is always wrong, divide the total number of male responses in that column into the number of males in that column, and then multiply the result by 100.

After we do this for each column, the final step is to compare across the column distributions within any one row. When we do this across the top row for the three samples in Table 2.4, we get the results shown in Table 2.5. We find that the difference in percentages between males and females is 14% for the first sample, zero for the second, and 38% for the third. This becomes the basis for our conclusion that the variables are unrelated in the second table and related in the third. Note that the 14% difference across the top row of the first table would reduce to 4% or increase to 24% by simply shifting one case into (or out of) the top left cell. This is our basis for rejecting the notion that meaningful conclusions can be drawn from this sample.

The point is that we cannot meaningfully compare the numbers of cases in each cell because the total number of males and females is different. We must first standardize the distributions by calculating the column percentages.

The preceding method of describing categorical data in two-way tables can also be represented visually in a clustered bar chart. Figure 2.8 shows such a chart, created using Excel, representing the third sample distribution in Table 2.5. Note the ease with which you can compare the males to females in each category of the "attitudes toward abortion" variable. Women are much more likely than men to see abortion as always wrong, and they are much less likely than men to see abortion as never wrong. Men's attitudes are more evenly distributed among the three response choices than are women's.

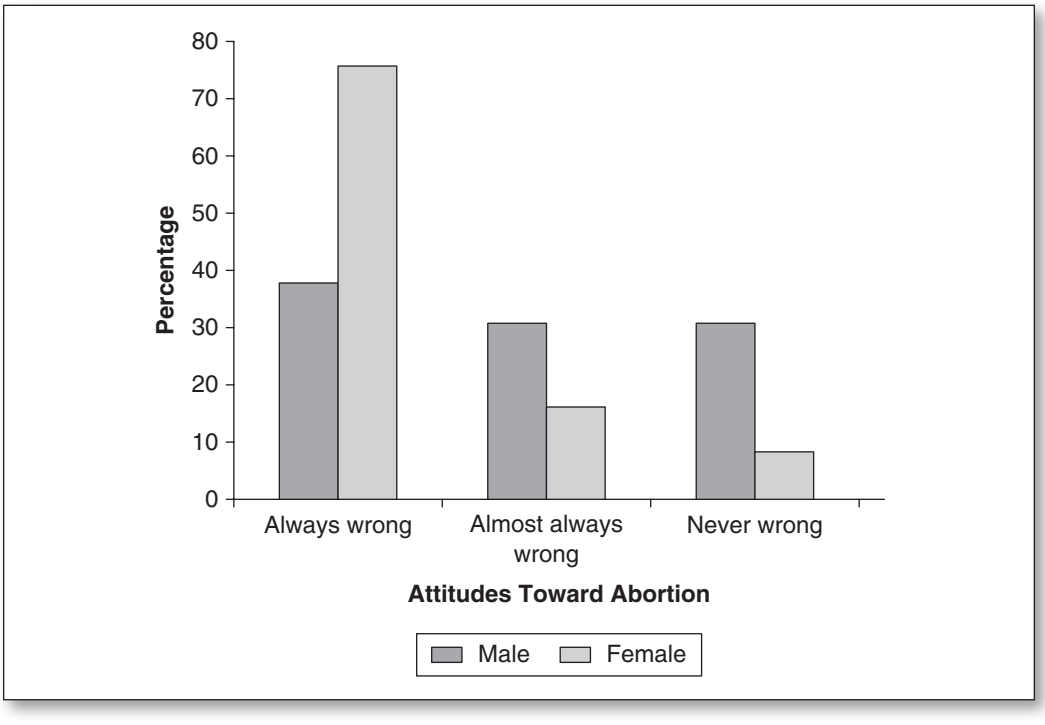
Table 2.5 Percentage Outcomes for Attitudes Toward Abortion, by Sex

	Sample 1		*Sample 2*		*Sample 3*	
Attitude Toward Abortion	*M*	*F*	*M*	*F*	*M*	*F*
Always wrong	40	54	76	76	38	76
Almost always wrong	30	32	14	14	31	16
Never wrong	30	14	10	10	31	8
Total	100	100	100	100	100	100
	(10)	(1,490)	(500)	(1,000)	(1,000)	(500)

SOURCE: Newton and Rudestam (1999).

NOTE: Numbers in parentheses are column marginal frequencies.

Figure 2.8 Clustered Bar Chart: Attitudes Toward Abortion by Sex



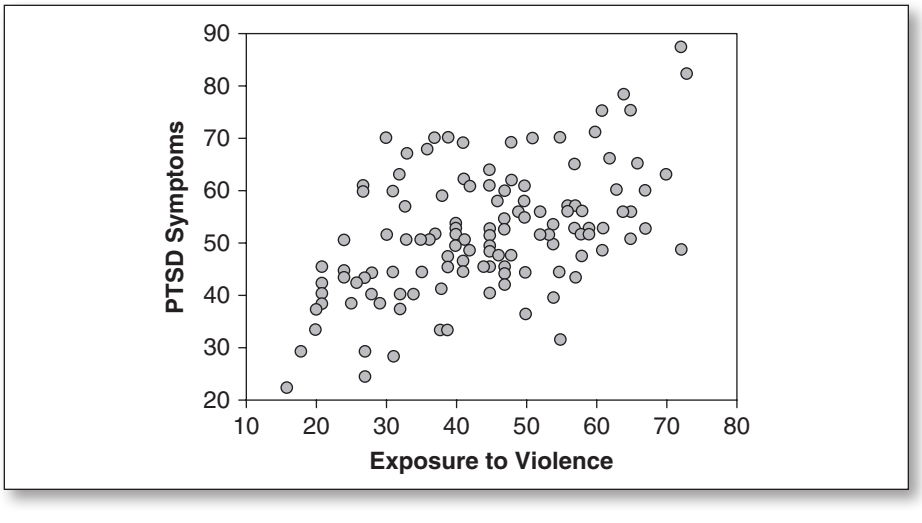
How Do I Use a Scatterplot to Examine Continuously Distributed Bivariate Data?

When working with continuously distributed data, the questions are the same as when working with categorical data: Is it reasonable to examine this relationship? If so, are the variables related?

If we are interested in examining the relationship between two continuously distributed variables, one way to obtain a preliminary understanding of this relationship is to plot the values of the variables on a graph, called a **scatterplot** or **scatter diagram.** Place one variable (the dependent or *Y* variable) on the vertical axis and the other (the independent or *X* variable) on the horizontal axis. Place appropriate scales to match the distribution of each variable on the axes and then place a dot on the graph to represent the location of each point. The example in Figure 2.9 shows the relationship between two scales, one assessing the level of exposure to violence and the other assessing symptoms of posttraumatic stress disorder (PTSD).

We can learn a number of things by inspection of the scatter diagram. First, we can determine if the relationship between the variables is positive or negative. **Positive relationships** represent relationships in which an increase in one variable is associated with an increase in a second. **Negative relationships** represent relationships in which an increase in one variable is associated with a

Figure 2.9 Scatterplot of the Relationship Between Exposure to Violence and Symptoms of Posttraumatic Stress Disorder (PTSD)



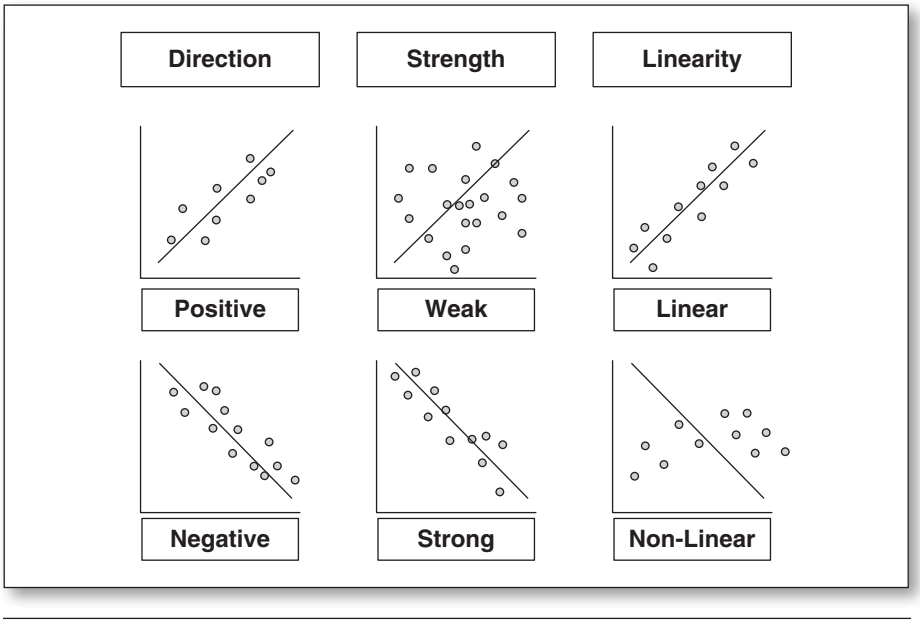
SOURCE: Newton and Rudestam (1999).

decrease in a second. Second, we can determine if the relationship is linear or nonlinear. **Linear relationships** are indicated when the pattern of dots on the scatter diagram appears to be straight, that is, when the points could be represented by drawing a straight line through them. Third, we can estimate the strength of the relationship between the variables. Strong relationships are indicated when the dots are very close to a straight line; weaker relationships are indicated by dots that are more scattered about, or farther from, a straight line.

For the example in Figure 2.9, the relationship is both positive and linear, as the pattern of dots goes from lower left to upper right in a "cigar-shaped" pattern that is typical of scatterplots representing linear positive relationships. The relationship also appears fairly strong. This judgment is supported by the value of the correlation coefficient of +.55. Figure 2.10 presents some examples of scatterplots with accompanying interpretations.

In conclusion, the first steps in data analysis begin with an examination of the data, first one variable at a time and then two variables at a time. These observations provide us with a picture of potential errors in data collection and entry and the distributions of the variables. We can then explore the data through the use of bivariate analyses, using crosstabulations and scatterplots.

Figure 2.10 Direction, Strength, and Linearity of Relationships as Shown by Scatterplots



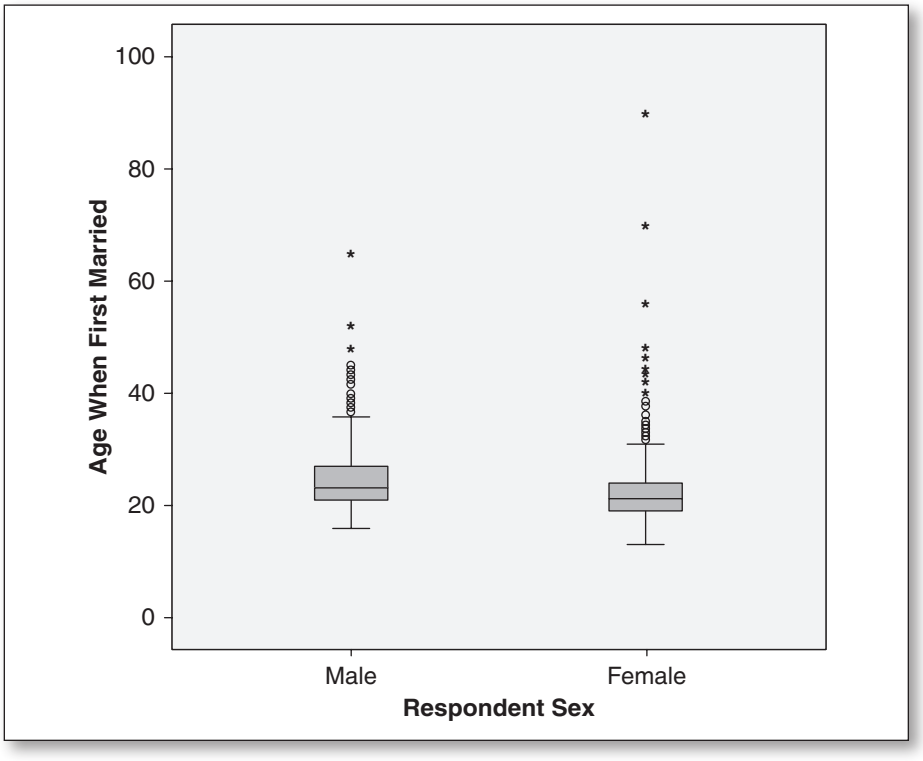
SOURCE: Newton and Rudestam (1999).

HOW DO I DISPLAY BIVARIATE DATA IF THE INDEPENDENT VARIABLE IS CATEGORICAL AND THE DEPENDENT VARIABLE CONTINUOUS?

So far we have discussed options for the display of bivariate data when both variables are categorical (crosstabulation) and both variables are continuous (scatterplots). What if the independent variable is categorical and the dependent variable continuous? We recommend two options for the basic display of data in this situation, side-by-side boxplots and clustered bar charts.

A side-by-side boxplot like the one in Figure 2.11 shows the distribution of each category of the independent variable (sex, in this example) in a separate boxplot with the scale of the dependent variable on the vertical or *Y*-axis. Display of age first married–by–sex distribution in this manner demonstrates a slightly older median age at first marriage for males (23) than for females (21)

Figure 2.11 Side-by-Side Boxplots of Age First Married by Sex

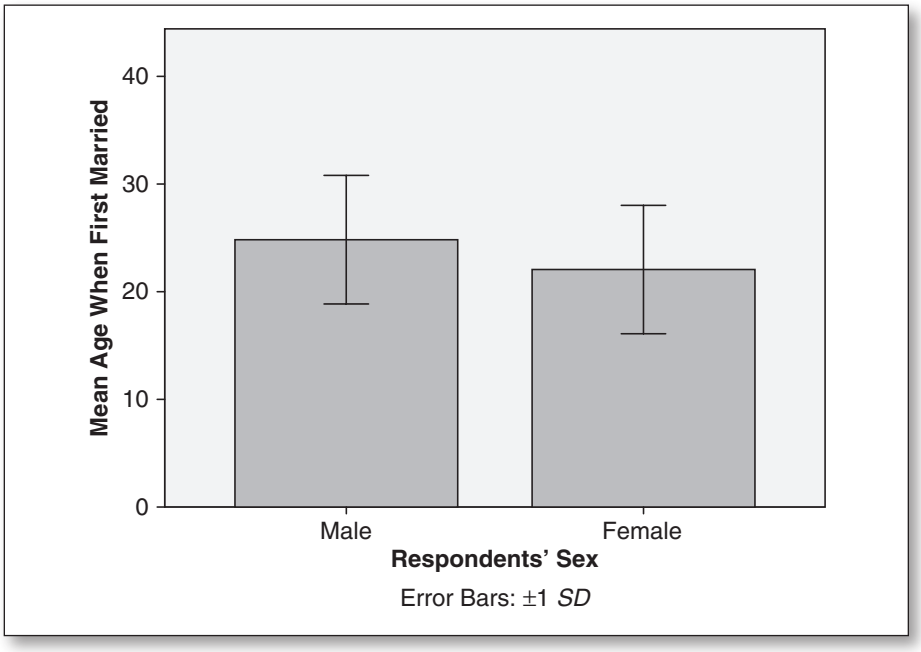


and a greater number of extreme outliers for females, including our one 90-year-old. One can also see that, despite the greater number and wider range of outliers for women, the middle 50%, represented by the box enclosing the data between the 25 and 75 percentiles, is both smaller and lower for women than for men.

Simple Bar Chart for Means

Another method for displaying the age first married–by–sex distribution, which emphasizes means and standard deviations as opposed to medians and quartiles and outliers, is to use a simple bar chart for means. In a bar chart for means, the bars show the mean levels, and the error bars represent 1 standard deviation on either side of the mean (a different set of error bars might show the standard error or confidence interval). Figure 2.12 makes it easy to see the older age at first marriage for men (24.84) compared to that for women (22.08), whereas the error bars, 1 standard deviation on either side of the mean, illustrate the fact that the standard deviations are approximately the same (5.98 for men and 5.97 for women).

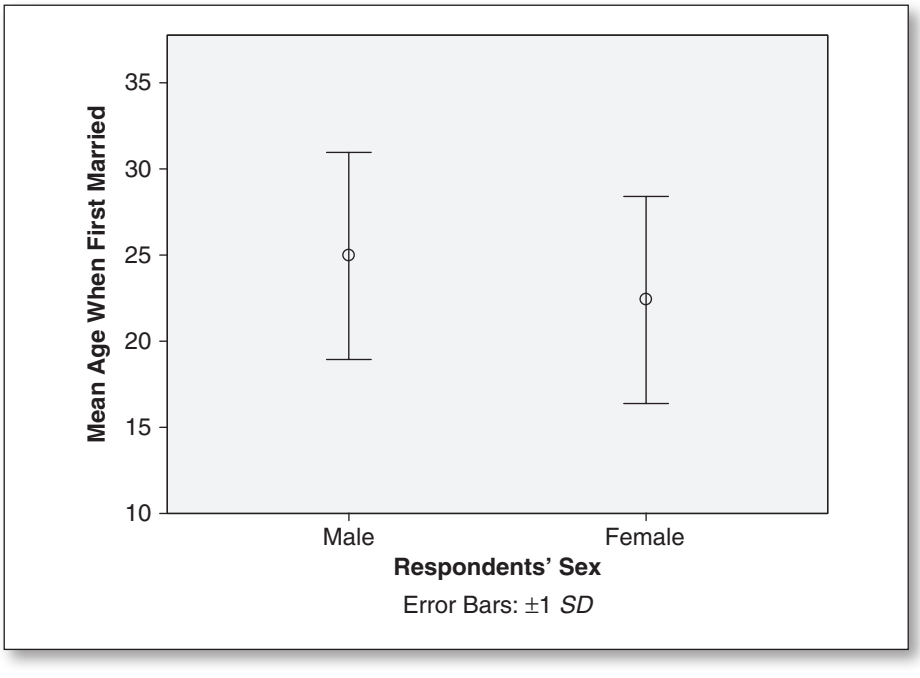
Figure 2.12 Bar Chart for Means with Error Bars Displaying Standard Deviations (bars represent means)



A similar chart for means, which represents the means with a dot rather than a bar, is shown in Figure 2.13.

Note that in Figure 2.13 we have emphasized the error bars by showing only the age range from 10 to 35, while in Figure 2.12 the age range was from 0 to 40 and the range of the side-by-side boxplots in Figure 2.11 was 0 to 100 (required to illustrate all of the outliers).

Figure 2.13 Bar Chart for Means with Error Bars Displaying Standard Deviations (dots represent means)



SUMMARY

The point of this chapter is to emphasize the importance of getting a sense of your data prior to moving forward with more complex analyses. This fairly straightforward process makes use of visual representations more than complex statistics, but through it you can detect errors in data entry, develop an awareness of the shape and structure of your data, and become aware of relationships that may exist between your variables.