Chapter 2

Characterizing data for analysis

2.1 Variables: their definition, classification, and use

In performing multivariate analysis, the investigator deals with numerous variables. In this chapter, we define what a variable is in Section 2.2. Section 2.3 presents a method of classifying variables that is sometimes useful in multivariate analysis since it allows one to check that a commonly used analysis has not been missed. Section 2.4 explains how variables are used in analysis and gives the common terminology for distinguishing between the two major uses of variables. Section 2.5 includes some examples of classifying variables and Section 2.6 discusses other characteristics of data and references exploratory data analysis.

2.2 Defining statistical variables

The word **variable** is used in statistically oriented literature to indicate a characteristic or property that is possible to measure. When we measure something, we make a numerical model of the thing being measured. We follow some rule for assigning a number to each level of the particular characteristic being measured. For example, the height of a person is a variable. We assign a numerical value to correspond to each person's height. Two people who are equally tall are assigned the same numeric value. On the other hand, two people of different heights are assigned two different values. Measurements of a variable gain their meaning from the fact that there exists unique correspondence between the assigned numbers and the levels of the property being measured. Thus two people with different assigned heights are not equally tall. Conversely, if a variable has the same assigned value for all individuals in a group, then this variable does not convey useful information to differentiate individuals in the group.

Physical measurements, such as height and weight, can be measured directly by using physical instruments. On the other hand, properties such as reasoning ability or the state of depression of a person must be measured indirectly. We might choose a particular intelligence test and define the variable

"intelligence" to be the score achieved on this test. Similarly, we may define the variable "depression" as the number of positive responses to a series of questions. Although what we wish to measure is the degree of depression, we end up with a count of yes answers to some questions. These examples point out a fundamental difference between direct physical measurements and abstract variables.

Often the question of how to measure a certain property can be perplexing. For example, if the property we wish to measure is the cost of keeping the air clean in a particular area, we may be able to come up with a reasonable estimate, although different analysts may produce different estimates. The problem becomes much more difficult if we wish to estimate the benefits of clean air.

On any given individual or thing we may measure several different characteristics. We would then be dealing with several variables, such as age, height, annual income, race, sex, and level of depression of a certain individual. Similarly, we can measure characteristics of a corporation, such as various financial measures. In this book we are concerned with analyzing data sets consisting of measurements on several variables for each individual in a given sample. We use the symbol P to denote the number of variables and the symbol N to denote the number of **individuals, observations, cases**, or **sampling units**.

2.3 Stevens's classification of variables

In the determination of the appropriate statistical analysis for a given set of data, it is useful to classify variables by type. One method for classifying variables is by the degree of sophistication evident in the way they are measured. For example, we can measure the height of people according to whether the top of their head exceeds a mark on the wall; if yes, they are tall; and if no, they are short. On the other hand, we can also measure height in centimeters or inches. The latter technique is a more sophisticated way of measuring height. As a scientific discipline advances, the measurement of the variables used in it tends to become more sophisticated.

Various attempts have been made to formalize variable classification. A commonly accepted system is that proposed by Stevens (1966). In this system, measurements are classified as **nominal, ordinal, interval,** or **ratio**. In deriving his classification, Stevens characterized each of the four types by a transformation that would not change a measurement's classification. In the subsections that follow, rather than discuss the mathematical details of these transformations, we present the practical implications for data analysis.

As with many classification schemes, Stevens's system is useful for some purposes but not for others. It should be used as a general guide to assist in characterizing the data and to make sure that a useful analysis is not overlooked. However, it should not be used as a rigid rule that ignores the purpose of the analysis or limits its scope (Velleman and Wilkinson, 1993).

Nominal variables

With **nominal variables** each observation belongs to one of several distinct categories. The categories are not necessarily numerical, although numbers may be used to represent them. For example, "sex" is a nominal variable. An individual's gender is either male or female. We may use any two symbols, such as M and F, to represent the two categories. In data analysis, numbers are used as the symbols since many computer programs are designed to handle only numerical symbols. Since the categories may be arranged in any desired order, any set of numbers can be used to represent them. For example, we may use 0 and 1 to represent males and females, respectively. We may also use 1 and 2 to avoid confusing zeros with blanks. Any two other numbers can be used as long as they are used consistently.

An investigator may rename the categories, thus performing a numerical operation. In doing so, the investigator must preserve the uniqueness of each category. Stevens expressed this last idea as a "basic empirical operation" that preserves the category to which the observation belongs. For example, two males must have the same value on the variable "sex," regardless of the two numbers chosen for the categories. Table 2.1 summarizes these ideas and presents further examples. Nominal variables with more than two categories, such as race or religion, may present special challenges to the multivariate data analyst. Some ways of dealing with these variables are presented in Chapter 9.

Ordinal variables

Categories are used for **ordinal variables** as well, but there also exists a known order among them. For example, in the Mohs Hardness Scale, minerals and rocks are classified according to ten levels of hardness. The hardest mineral is diamond and the softest is talc (Pough, 1996). Any ten numbers can be used to represent the categories, as long as they are ordered in magnitude. For instance, the integers 1–10 would be natural to use. On the other hand, any sequence of increasing numbers may also be used. Thus, the basic empirical operation defining ordinal variables is whether one observation is greater than another. For example, we must be able to determine whether one mineral is harder than another. Hardness can be tested easily by noting which mineral can scratch the other. Note that for most ordinal variables there is an underlying continuum being approximated by artificial categories. For example, in the above hardness scale fluorite is defined as having a hardness of 4, and calcite, 3. However, there is a range of hardness between these two numbers not accounted for by the scale.

Table 2.1: ***Stevens's measurement system***

Type of measurement	Basic empirical operation	Examples
Nominal	Determine equality of categories	Company names Race Religion Soccer players' numbers
Ordinal	Determine greater than or less than (ranking)	Hardness of minerals Socioeconomic status Rankings of wines
Interval	Determine equality of differences between levels	Temperature in degrees Fahrenheit Calendar dates
Ratio	Determine equality of ratios of levels	Height Weight Density Difference in time

Often investigators classify people, or ask them to classify themselves, along some continuum (see Luce and Narens, 1987). For example, a physician may classify a patient's disease status as none = 1, mild = 2, moderate = 3, and severe = 4. Clearly, increasing numbers indicate increasing severity, but it is not certain that the difference between not having an illness and having a mild case is the same as between having a mild case and a moderate case. Hence, according to Stevens's classification system, this is an ordinal variable.

Interval variables

An **interval variable** is a variable in which the differences between successive values are always the same. For example, the variable "temperature," in degrees Fahrenheit, is measured on the interval scale since the difference between 12° and 13° is the same as the difference between 13° and 14° or the difference between any two successive temperatures. In contrast, the Mohs Hardness Scale does not satisfy this condition since the intervals between successive categories are not necessarily the same. The scale must satisfy the basic empirical operation of preserving the equality of intervals.

Ratio variables

Ratio variables are interval variables with a natural point representing the origin of measurement, i.e., a natural zero point. For instance, height is a ratio variable since zero height is a naturally defined point on the scale. We may change the unit of measurement (e.g., centimeters to inches), but we would still preserve the zero point and also the ratio of any two values of height. Temperature is not a ratio variable since we may choose the zero point arbitrarily, thus not preserving ratios.

There is an interesting relationship between interval and ratio variables. The difference between two interval variables is a ratio variable. For example, although time of day is measured on the interval scale, the length of a time period is a ratio variable since it has a natural zero point.

Other classifications

Other methods of classifying variables have also been proposed. Many authors use the term **categorical** to refer to nominal and ordinal variables where categories are used.

We mention, in addition, that variables may be classified as discrete or continuous. A variable is called **continuous** if it can take on any value in a specified range. Thus the height of an individual may be 70 or 70.4539 inches. Any numerical value in a certain range is a conceivable height.

A variable that is not continuous is called **discrete**. A discrete variable may take on only certain specified values. For example, counts are discrete variables since only zero or positive integers are allowed. In fact, all nominal and ordinal variables are discrete. Interval and ratio variables can be continuous or discrete. This latter classification carries over to the possible distributions assumed in the analysis. For instance, the normal distribution is often used to describe the distribution of continuous variables.

Statistical analyses have been developed for various types of variables. In Chapter 5 a guide to selecting the appropriate descriptive measures and multivariate analyses will be presented. The choice depends on how the variables are used in the analysis, a topic that is discussed next.

2.4 How variables are used in data analysis

The type of data analysis required in a specific situation is also related to the way in which each variable in the data set is used. Variables may be used to measure outcomes or to explain why a particular outcome resulted. For example, in the treatment of a given disease a specific drug may be used. The **outcome variable** may be a discrete variable classified as "cured" or "not cured." The outcome variable may depend on several characteristics of the patient such as age, genetic background, and severity of the disease. These characteristics are sometimes called **explanatory** or **predictor variables**. Equivalently, we may call the outcome the **dependent variable** and the characteristics the **independent variable**. The latter terminology is very common in statistical litera-

ture. This choice of terminology is unfortunate in that the "independent" variables do not have to be statistically independent of each other. Indeed, these independent variables are usually interrelated in a complex way. Another disadvantage of this terminology is that the common connotation of the words implies a causal model, an assumption not needed for the multivariate analyses described in this book. In spite of these drawbacks, the widespread use of these terms forces us to adopt them.

In other situations the dependent or outcome variable may be treated as a continuous variable. For example, in household survey data we may wish to relate monthly expenditure on cosmetics per household to several explanatory or independent variables such as the number of individuals in the household, their gender, and the household income.

In some situations the roles that the various variables play are not obvious and may also change, depending on the question being addressed. Thus a data set for a certain group of people may contain observations on their sex, age, diet, weight, and blood pressure. In one analysis, we may use weight as a dependent or outcome variable with height, sex, age, and diet as the independent or predictor variables. In another analysis, blood pressure might be the dependent or outcome variable, with weight and other variables considered as independent or predictor variables.

In certain exploratory analyses all the variables may be used as one set with no regard to whether they are dependent or independent. For example, in the social sciences a large number of variables may be defined initially, followed by attempts to combine them into a smaller number of summary variables. In such an analysis the original variables are not classified as dependent or independent. The summary variables may later be used either as outcome or predictor variables. In Chapter 5 multivariate analyses described in this book will be characterized by the situations in which they apply according to the types of variables analyzed and the roles they play in the analysis.

2.5 Examples of classifying variables

In the depression data example several variables are measured on the nominal scale: sex, marital status, employment, and religion. The general health scale is an example of an ordinal variable. Income and age are both ratio variables. No interval variable is included in the data set. A partial listing and a codebook for this data set are given in Chapter 3.

One of the questions that may be addressed in analyzing these data is "Which factors are related to the degree of psychological depression of a person?" The variable "cases" may be used as the dependent or outcome variable since an individual is considered a case if his or her score on the depression scale exceeds a certain level. "Cases" is an ordinal variable, although it can be considered nominal because it has only two categories. The independent

or predictor variable could be any or all of the other variables (except ID and measures of depression). Examples of analyses without regard to variable roles are given in Chapters 14 and 15 using the variables C_1 to C_{20} in an attempt to summarize them into a small number of components or factors.

Sometimes, Stevens's classification system is difficult to apply, and two investigators could disagree on a given variable. For example, there may be disagreement about the ordering of the categories of a socioeconomic status variable. Thus the status of blue-collar occupations with respect to the status of certain white-collar occupations might change over time or from culture to culture. So such a variable might be difficult to justify as an ordinal variable, but we would be throwing away valuable information if we used it as a nominal variable. Despite these difficulties, Stevens's system is useful in making decisions on appropriate statistical analysis, as will be discussed in Chapter 5.

2.6 Other characteristics of data

Data are often characterized by whether the measurements are accurately taken and are relatively error free, and by whether they meet the assumptions that were used in deriving statistical tests and confidence intervals. Often, an investigator knows that some of the variables are likely to have observations that have errors. If the effect of an error causes the numerical value of an observation to not be in line with the numerical values of most of the other observations, these extreme values may be called **outliers** and should be considered for removal from the analysis. But other observations may not be accurate and still be within the range of most of the observations. Data sets that contain a sizeable portion of inaccurate data or errors are called "dirty" data sets.

Special statistical methods have been developed that are resistant to the effects of dirty data. Other statistical methods, called robust methods, are insensitive to departures from underlying model assumptions. In this book, we do not present these methods but discuss finding outliers and give methods of determining if the data meet the assumptions. For further information on statistical methods that are well suited for dirty data or require few assumptions, see Hoaglin *et al.* (1985), Schwaiger and Opitz (2003), or Fox and Long (1990).

2.7 Summary

In this chapter statistical variables were defined. Their types and the roles they play in data analysis were discussed. Stevens's classification system was described.

These concepts can affect the choice of analyses to be performed, as will be discussed in Chapter 5.

2.8 Problems

2.1 Classify the following types of data by using Stevens's measurement system: decibels of noise level, father's occupation, parts per million of an impurity in water, density of a piece of bone, rating of a wine by one judge, net profit of a firm, and score on an aptitude test.

2.2 In a survey of users of a walk-in mental health clinic, data have been obtained on sex, age, household roster, race, education level (number of years in school), family income, reason for coming to the clinic, symptoms, and scores on screening examination. The investigator wishes to determine what variables affect whether or not coercion by the family, friends, or a governmental agency was used to get the patient to the clinic. Classify the data according to Stevens's measurement system. What would you consider to be possible independent variables? Dependent variables? Do you expect the dependent variables to be independent of each other?

2.3 For the chronic respiratory study data described in Appendix A, classify each variable according to Stevens's scale and according to whether it is discrete or continuous. Pose two possible research questions and decide on the appropriate dependent and independent variables.

2.4 Repeat problem 2.3 for the lung cancer data set described in Table 13.1.

2.5 From a field of statistical application (perhaps your own field of specialty), describe a data set and repeat the procedures described in Problem 2.3.

2.6 If the RELIG variable described in Table 3.4 of this text was recoded 1 = Catholic, 2 = Protestant, 3 = Jewish, 4 = none, and 5 = other, would this meet the basic empirical operation as defined by Stevens for an ordinal variable?

2.7 Give an example of nominal, ordinal, interval, and ratio variables from a field of application you are familiar with.

2.8 Data that are ordinal are often analyzed by methods that Stevens reserved for interval data. Give reasons why thoughtful investigators often do this.

2.9 The Parental HIV data set described in Appendix A includes the following variables: job status of mother (JOBMO, 1=employed, 2=unemployed, and 3=retired/disabled) and mother's education (EDUMO, 1=did not complete high school, 2=high school diploma/GED, and 3=more than high school). Classify these two variables using Stevens's measurement system.

2.10 Give an example from a field that you are familiar with of an increased sophistication of measuring that has resulted in a measurement that used to be ordinal now being interval.

Chapter 3

Preparing for data analysis

3.1 Processing data so they can be analyzed

Once the data are available from a study there are still a number of steps that must be undertaken to get them into shape for analysis. This is particularly true when multivariate analyses are planned since these analyses are often done on large data sets. In this chapter we provide information on topics related to data processing.

Section 3.2 describes the statistical software packages used in this book. Note that several other statistical packages offer an extensive selection of multivariate analyses. In addition, almost all statistical packages and even some of the spreadsheet programs include at least multiple regression as an option.

The next topic discussed is data entry (Section 3.3). Survey data collection is performed more and more using computers directly via Computer Assisted Personal Interviewing (CAPI), Audio Computer Assisted Self Interviewing (ACASI), or via the Internet. For example, SurveyMonkey is a commercially available program that facilitates sending and collecting surveys via the Internet. Nonetheless, paper and pencil interviews or mailed questionnaires are still a major form of data collection. The methods that need to be used to enter the information obtained from paper and pencil interviews into a computer depend on the size of the data set. For a small data set there are a variety of options since cost and efficiency are not important factors. Also, in that case the data can be easily screened for errors simply by visual inspection. But for large data sets, careful planning of data entry is necessary since costs are an important consideration along with getting an error-free data set available for analysis. Here we summarize the data input options available in the statistical software packages used in this book and discuss the useful options.

Section 3.4 covers combining and updating data sets. The operations used and the options available in the various packages are described. Initial discussion of missing values, outliers, and transformations is given and the need to save results is stressed. Finally, in Section 3.5 we introduce a multivariate data set that will be widely used in this book and summarize the data in a codebook.

We want to stress that the procedures discussed in this chapter can be time

consuming and frustrating to perform when large data sets are involved. Often the amount of time used for data entry, editing, and screening can far exceed that used on statistical analysis. It is very helpful to either have computer expertise yourself or have access to someone you can get advice from occasionally.

3.2 Choice of a statistical package

There is a wide choice of statistical software packages available. Many of these packages, however, are quite specialized and do not include many of the multivariate analyses given in this book. For example, there are statistical packages that are aimed at particular areas of application or give tests for exact statistics that are more useful for other types of work. In choosing a package for multivariate analysis, we recommend that you consider the statistical analyses listed in Table 5.2 and check whether the package includes them.

In some cases the statistical package is sold as a single unit and in others you purchase a basic package, but you have a choice of additional programs so you can buy what you need. Some programs require yearly license fees.

Ease of use

Some packages are easier to use than others, although many of us find this difficult to judge–we like what we are familiar with. In general, the packages that are simplest to use have two characteristics. First, they have fewer options to choose from and these options are provided automatically by the program with little need for programming by the user. Second, they use the "point and click" method of choosing what is done rather than writing out statements. The point and click method is even simpler to learn if the package uses options similar to ones found in word processing, spreadsheet, or database management packages. But many current point and click programs do not leave the user with an audit trail of what choices have been made.

On the other hand, software programs with extensive options have obvious advantages. Also, the use of written statements (or *commands*) allows you to have a written record of what you have done. This record can be particularly useful in large-scale data analyses that extend over a considerable period of time and involve numerous investigators. Still other programs provide the user with a programming language that allows the users great freedom in what output they can obtain.

Packages used in this book

In this book, we make specific reference to six general-purpose statistical software packages: R, S-PLUS, SAS, SPSS, Stata, and STATISTICA, listed in alphabetical order. Although S-PLUS has been renamed TIBCO Spotfire S+,

we continue to use the name S-PLUS since it is more familiar to many readers. For this edition, we also include R, a popular package that is being increasingly used by statisticians. It is a freeware, i.e., it consists of software that is available free of charge. It offers a large number of multivariate analyses, including most of the ones discussed in this book. Many of the commands in R are similar to those in S-PLUS and therefore we discuss them together. The software versions we use in this book are those available at the end of December 2010.

S-PLUS and R can be used on quite different levels. The simplest is to access it through Microsoft Excel and then run the programs using the usual point and click operations. The most commonly used analyses can be performed this way. Alternatively, S-PLUS and R can be used as a language created for performing statistical analyses. The user writes the language expressions that are read and immediately executed by the program. This process allows the user to write a function, run it, see what happens, and then use the result of the function in a second function. Effort is required to learn the language but, similar to SAS and Stata, it provides the user with a highly versatile programming tool for statistical computing. There are numerous books written on writing programs in S-PLUS and R for different areas of application; for example, see Braun and Murdoch (2007), Crawley (2002), Dalgaard (2008), Everitt (2007), Hamilton (2009) or Heiberger and Neuwirth (2009).

The SAS philosophy is that the user should string together a sequence of procedures to perform the desired analysis. SAS provides a lot of versatility to the user since the software provides numerous possible procedures. It also provides extensive capability for data manipulations. Effort is required to learn the language, but it provides the user with a highly versatile programming tool for statistical computing. Some data management and analysis features are available via point and click operations (SAS/LAB and SAS/ASSIST). Numerous texts have been written on using SAS; for example, see Khattree and Naik (2003), Der and Everitt (2008), or Freund and Littell (2006).

SPSS was originally written for survey applications. The manuals are easy to read and the available options are well chosen. It offers a number of comprehensive programs and users can choose specific options that they desire. It provides excellent data entry programs and data manipulation procedures. It can be used either with point and click or command modes. In addition to the manuals, books such as the ones by Abu-Bader (2010) or Gerber and Finn (2005) are available.

Stata is similar to SAS in that an analysis consists of a sequence of commands with their own options. Analyses can also be performed via a point and click environment. Features are available to easily log and rerun an analysis. Stata is relatively inexpensive and contains features that are not found in other programs. It also includes numerous data management features and a very rich set of graphics options. Several books are available which discuss

statistical analysis using Stata; see Rabe-Hesketh and Everitt (2007), Hills and De Stavola (2009), or Cleves, Gould and Gutierrez (2008).

STATISTICA is a very easy software to use. It runs using the usual Windows point and click operations. It has a comprehensive list of options for multivariate analysis. STATISTICA 6 was used in this text. It allows the user to easily record logs of analyses so that routine jobs can be run the same way in the future. The user can also develop custom applications. See the statsoft.com web site for a wide listing of books on using STATISTICA.

When you are learning to use a package for the first time, there is no substitute for reading the on-line HELP, manuals, or texts that present examples. However, at times the sheer number of options presented in these programs may seem confusing, and advice from an experienced user may save you time. Many programs offer default options, and it often helps to use these when you run a program for the first time. In this book, we frequently recommend which options to use. On-line HELP is especially useful when it is programmed to offer information needed for the part of the program you are currently using (context sensitive). Links to the preceding software programs can be found in the UCLA web site cited in the preface.

There are numerous statistical packages that are not included in this book. We have tried to choose those that offer a wide range of multivariate techniques. This is a volatile area with new packages being offered by numerous software firms.

For information on other packages, you can refer to the statistical computing software review sections of *The American Statistician, PC Magazine*, or journals in your own field of interest.

3.3 Techniques for data entry

Appropriate techniques for entering data for analysis depend mainly on the size of the data set and the form in which the data set is stored. As discussed below, all statistical packages use data in a spreadsheet (or rectangular) format. Each column represents a specific variable and each row has the data record for a case or observation. The variables are in the same order for each case. For example, for the depression data set given later in this chapter, looking only at the first three variables and four cases, we have

ID	Sex	Age
1	2	68
2	1	58
3	2	45
4	2	50

where for the variable "sex," 1 = male and 2 = female, and "age" is given in years.

Normally each row represents an individual case. What is needed in each row depends on the unit of analysis for the study. By unit of analysis, we mean what is being studied in the analysis. If the individual is the unit of analysis, as it usually is, then the data set just given is in a form suitable for analysis. Another situation is when the individuals belong to one household, and the unit of analysis is the household but data have been obtained from several individuals in the household. Alternatively, for a company, the unit of analysis may be a sales district and sales made by different salespersons in each district are recorded. Data sets given in the last two examples are called hierarchical data sets and their form can get to be quite complex. Some statistical packages have limited capacity to handle hierarchical data sets. In other cases, the investigator may have to use a relational database package such as Access to first get the data set into the rectangular or spreadsheet form used in the statistical package.

As discussed below, either one or two steps are involved in data entry. The first one is actually entering the data into the computer. If the data are not entered directly into the statistical package being used, a second step of transferring the data to the desired statistical package must be performed.

Data entry

Before entering the actual data in most statistical, spreadsheet, or database management packages, the investigator first names the file where the data are stored, states how many variables will be entered, names the variables, and provides information on these variables. Note that in the example just given we listed three variables which were named for easy use later. The file could be called "depress." Statistical packages commonly allow the user to designate the format and type of variable, e.g., numeric or alphabetic, calendar date, or categorical. They allow you to specify missing value codes, the length of each variable, and the placement of the decimal points. Each program has slightly different features so it is critical to read the appropriate online HELP statements or manual, particularly if a large data set is being entered.

The two commonly used formats for data entry are the **spreadsheet** and the **form**. By spreadsheet, we mean the format given previously where the columns are the variables and the rows the cases. This method of entry allows the user to see the input from previous records, which often gives useful clues if an error in entry is made. The spreadsheet method is very commonly used, particularly when all the variables can be seen on the screen without scrolling. Many persons doing data entry are familiar with this method due to their experience with spreadsheet programs, so they prefer it.

With the form method, only one record, the one being currently entered, is on view on the screen. There are several reasons for using the form method. An entry form can be made to look like the original data collection form so that the data entry person sees data in the same place on the screen as it is

in the collection form. A large number of variables for each case can be seen on a computer monitor screen and they can be arranged in a two-dimensional array, instead of just the one-dimensional array available for each case in the spreadsheet format. Flipping pages (screens) in a display may be simpler than scrolling left or right for data entry. Short coding comments can be included on the screen to assist in data entry. Also, if the data set includes alphabetical information such as short answers to open-ended questions, then the form method is preferred.

The choice between these two formats is largely a matter of personal preference, but in general the spreadsheet is used for data sets with a small or medium number of variables and the form is used for a larger number of variables. In some cases a scanner can be used to enter the data and then an optical character recognition program converts the image to the desired text and numbers.

To make the discussion more concrete, we present the features given in a specific data entry package. The SPSS data entry program provides a good mix of features that are useful in entering large data sets. It allows either spreadsheet or form entry and switching back and forth between the two modes. In addition to the features already mentioned, SPSS provides what is called "skip and fill." In medical studies and surveys, it is common that if the answer to a certain question is no, a series of additional questions can then be skipped. For example, subjects might be asked if they ever smoked, and if the answer is yes they are asked a series of questions on smoking history. But if they answer no, these questions are not asked and the interviewer skips to the next section of the interview. The skip-and-fill option allows the investigator to specify that if a person answers no, the smoking history questions are automatically filled in with specified values and the entry cursor moves to the start of the next section. This saves a lot of entry time and possible errors.

Another feature available in many packages is range checking. Here the investigator can enter upper and lower values for each variable. If the data entry person enters a value that is either lower than the low value or higher than the high value, the data entry program provides a warning. For example, for the variable "sex," if an investigator specifies 1 and 2 as possible values and the data entry person hits a 3 by mistake, the program issues a warning. This feature, along with input by forms or spreadsheet, is available also in SAS.

Each software has its own set of features and the reader is encouraged to examine them before entering medium or large data sets, to take advantage of them.

Mechanisms of entering data

Data can be entered for statistical computation from different sources. We will discuss four of them.

1. entering the data along with the program or procedure statements for a batch-process run;
2. using the data entry features of the statistical package you intend to use;
3. entering the data from an outside file which is constructed without the use of the statistical package;
4. importing the data from another package using the operating system such as Windows or MAC OS.

The first method can only be used with a limited number of programs which use program or procedure statements, for example R, S-PLUS or SAS. It is only recommended for very small data sets that are not going to be used very many times. For example, a SAS data set called "depress" could be made by stating:

```
data depress;
  input id sex age;
  cards;
1   2   68
2   1   58
3   2   45
4   2   50
:
run;
```

Similar types of statements can be used for the other programs which use the spreadsheet type of format.

The disadvantage of this type of data entry is that there are only limited editing features available to the person entering the data. No checks are made as to whether or not the data are within reasonable ranges for this data set. For example, all respondents were supposed to be 18 years old or older, but there is no automatic check to verify that the age of the third person, who was 45 years old, was not erroneously entered as 15 years. Another disadvantage is that the data set disappears after the program is run unless additional statements are made. In small data sets, the ability to save the data set, edit typing, and have range checks performed is not as important as in larger data sets.

The second strategy is to use the data entry package or system provided by the statistical program you wish to use. This is always a safe choice as it means that the data set is in the form required by the program and no data transfer problems will arise. Table 3.1 summarizes the built-in data entry features of the six statistical packages used in this book. Note that for SAS, Proc COMPARE can be used to verify the data after they are entered. In general, as can be seen in Table 3.1, SPSS and SAS have extensive data entry features.

The third method is to use another statistical software package, data entry package, word processor, spreadsheet, or data management program to enter

Table 3.1: ***Built-in data entry features of the statistical packages***

	S-PLUS/R	SAS	SPSS	Stata	STATISTICA
Spreadsheet entry	Yes	Yes	Yes	Yes	Yes
Form entry	No	Yes	Yes	No	No
Range check	User	Yes	Yes	No	No
Logical check	User	Yes	Yes	No	No
Skip and fill	User	Use SCL	Yes	No	No
Verify mode	No	No	Yes	No	No

the data into a spreadsheet format. The advantage of this method is that an available program that you are familiar with can be used to enter the data. Two commonly used programs for data entry are Excel and Access. Excel provides entry in the form of a spreadsheet and is widely available. Access allows entry using forms and provides the ability to combine different data sets. Once the data sets are combined in Access, it is straightforward to transfer them to Excel. Many of the statistical software packages import Excel files. In addition, many of the statistical packages allow the user to import data from other statistical packages. For example, R and S-PLUS will import SAS, SPSS, and Stata data files. One suggestion is to first check the manual or HELP for the statistical package you wish to use to see which types of data files it can import.

A widely used transfer method is to create an ASCII file from the data set. ASCII (American Standard Code for Information Interchange) files can be created by almost any spreadsheet, data management, or word processing program. Instructions for reading ASCII files are given in the statistical packages. The disadvantage of transferring ASCII files is that typically only the data are transferred, and variable names and information concerning the variables have to be reentered into the statistical package. This is a minor problem if there are not too many variables. If this process appears to be difficult, or if the investigator wishes to retain the variable names, then they can run a special-purpose program such as STAT/TRANSFER, DBMS/COPY (available from Data Flux Corporation) or DATA JUNCTION that will copy data files created by a wide range of programs and put them into the right format for access by any of a wide range of statistical packages.

Finally, if the data entry program and the statistical package both use the Windows operating system, then three methods of transferring data may be considered depending on what is implemented in the programs. First, the data in the data entry program may be highlighted and moved to the statistical package using the usual copy and paste options. Second, dynamic data exchange

Table 3.2: ***Data management features of the statistical packages***

	S-PLUS/R	SAS	SPSS	Stata	STATISTICA
Merging data sets	merge	MERGE statement	MATCH FILES	merge	merge
Adding data sets	rbind cbind	PROC APPEND or set statement	ADD FILES	append	add cases add variables
Hierarchical data sets	User written functions	Write multiple output statements: RETAIN	CASESTOVARS	reshape	stacking
Importing data (types)	ASCII, spreadsheets, databases stat-packages	ASCII, ACCESS: spreadsheets, databases	ASCII, spreadsheets, databases	ASCII, spreadsheets, databases	ASCII, spreadsheets, databases
Exporting data (types)	ASCII, spreadsheet, databases	ASCII, ACCESS: spreadsheets, databases	ASCII, spreadsheets, databases	ASCII, spreadsheets, databases	ASCII, spreadsheets, databases
Calender dates	Yes	Yes	Yes	Yes	Yes
Transpose data	t	PROC TRANSPOSE	FLIP	xpose	Transpose
Range limit checks	Yes	Yes	Yes	Yes	Yes
Missing value imputation	Yes	MI and MIANALYZE	MULTIPLE IMPUTATION	mi	Mean substitution

(DDE) can be used to transfer data. Here the data set in the statistical package is dynamically linked to the data set in the entry program. If you correct a variable for a particular case in the entry program, the identical change is made in the data set in the statistical package, Third, object linking and embedding (OLE) can be used to share data between a program used for data entry and statistical analysis. Here also the data entry program can be used to edit the data in the statistical program. The investigator can activate the data entry program from within the statistical package program. MAC users often find it simplest to enter their data in Excel and then transfer their Excel file to Windows-based computers for analysis. Transfers can be made by disk or by a third party software package called DAVE.

If you have a very large data set to enter, it is often sensible to use a professional data entering service. A good service can be very fast and can offer different levels of data checking and advice on which data entry method to use. But whether or not a professional service is used, the following suggestions may be helpful for data entry.

1. Whenever possible, code information in numbers not letters.
2. Code information in the most detailed form you will ever need. You can use the statistical program to aggregate the data into coarser groupings later. For example, it is better to record age as the exact age at the last birthday rather than to record the ten-year age interval into which it falls.
3. The use of range checks or maximum and minimum values can eliminate the entry of extreme values but they do not guard against an entry error that falls within the range. If minimizing errors is crucial then the data can be entered twice into separate data files. One data file can be subtracted from the other and the resulting nonzeros examined. Alternatively, some data entry programs have a verify mode where the user is warned if the first entry does not agree with the second one (SPSS).
4. If the data are stored on a personal computer, then backup copies should be made on an external storage device such as a CD or DVD. Backups should be updated regularly as changes are made in the data set. Particularly when using Windows programs, if dynamic linking is possible between analysis output and the data set, it is critical to keep an unaltered data set.
5. For each variable, use a code to indicate missing values. The various programs each have their own way to indicate missing values. The manuals or HELP statements should be consulted so that you can match what they require with what you do.

To summarize, there are three important considerations in data entry: accuracy, cost, and ease of use of the data file. Whichever system is used, the investigator should ensure that the data file is free of typing errors, that time and money are not wasted, and that the data file is readily available for future data management and statistical analysis.

3.4 Organizing the data

Prior to statistical analysis, it is often necessary to make some changes in the data set. Table 3.2 summarizes the common options in the programs described in this book.

Combining data sets

Combining data sets is an operation that is commonly performed. For example, in biomedical studies, data may be taken from medical history forms, a questionnaire, and laboratory results for each patient. These data for a group of patients need to be combined into a single rectangular data set where the rows are the different patients and the columns are the combined history, questionnaire, and laboratory variables. In longitudinal studies of voting intentions, the questionnaire results for each respondent must be combined across time periods in order to analyze change in voting intentions of an individual over time. There are essentially two steps in this operation. The first is sorting on some key variable (given different names in different packages) which must be included in the separate data sets to be merged. Usually this key variable is an identification or ID variable (case number). The second step is combining the separate data sets side-by-side, matching the correct records with the correct person using the key variable. Sometimes one or more of the data items are missing for an individual. For example, in a longitudinal study it may not be possible to locate a respondent for one or more of the interviews. In such a case, a symbol or symbols indicating missing values will be inserted into the spaces for the missing data items by the program. This is done so that you will end up with a rectangular data set or file, in which information for an individual is put into the proper row, and missing data are so identified.

Data sets can be combined in the manner described above in SAS by using the MERGE statement followed by a BY statement and the variable(s) to be used to match the records. (The data must first be sorted by the values of the matching variable, say ID.) An UPDATE statement can also be used to add variables to a master file. In SPSS, you simply use the JOIN MATCH command followed by the data files to be merged if you are certain that the cases are already listed in precisely the same order and each case is present in all the data files. Otherwise, you first sort the separate data files on the key variable and use the JOIN MATCH command followed by the BY key variable. In Stata, you USE the first data file and then use a MERGE key variable USING the second data file statement. STATISTICA has a merge function and S-PLUS also has the CBIND function or a merge BY.X or BY.Y argument can be used for more

complex situations (see their help file). This step can also be done using a cut and paste operation in many programs.

In any case, it is highly desirable to list the data set to determine that the merging was done in the manner that you intended. If the data set is large, then only the first and last 25 or so cases need to be listed to see that the results are correct. If the separate data sets are expected to have missing values, you need to list sufficient cases so you can see that missing records are correctly handled.

Another common way of combining data sets is to put one data set at the end of another data set or to interleave the cases together based on some key variable. For example, an investigator may have data sets that are collected at different places and then combined together. In an education study, student records could be combined from two high schools, with one simply placed at the bottom of the other set. This is done by using the Proc APPEND in SAS. In SPSS the JOIN command with the keyword ADD can be used to combine cases from two to five data files or with specification of a key variable, to interleave. In Stata the APPEND command is used and in S-PLUS the RBIND function. In STATISTICA, the MERGE procedure is used. This step can also be done as a cut and paste operation in many programs.

It is also possible to update the data files with later information using the editing functions of the package. Thus a single data file can be obtained that contains the latest information, if this is desired for analysis. This option can also be used to replace data that were originally entered incorrectly.

When using a statistical package that does not have provision for merging data sets, it is recommended that a spreadsheet program be used to perform the merging and then, after a rectangular data file is obtained, the resulting data file can be transferred to the desired statistical package. In general, the newer spreadsheet programs have excellent facilities for combining data sets side-by-side or for adding new cases.

Missing values

There are two types of missing data. The first type occurs when no information is obtained from a case, individual, or sampling unit. This type is called **unit nonresponse**. For example, in a survey it may be impossible to reach a potential respondent or the subject may refuse to answer. In a biomedical study, records may be lost or a laboratory animal may die of unrelated causes prior to measuring the outcome. The second type of nonresponse occurs when the case, individual, or sampling unit is available but yields incomplete information. For example, in a survey the respondent may refuse to answer questions on income or only fill out the first page of a questionnaire. Busy physicians may not completely fill in a medical record. This type of nonresponse is called **item nonresponse**. In general, the more control the investigator has of the sampling

units, the less apt unit nonresponse or item nonresponse is to occur. In surveys the investigator often has little or no control over the respondent, so both types of nonresponse are apt to happen. For this reason, much of the research on handling nonresponse has been done in the survey field and the terminology used reflects this emphasis.

The seriousness of either unit nonresponse or item nonresponse depends mainly on the magnitude of the nonresponse and on the characteristics of the nonresponders. If the proportion of nonresponse is very small, it is seldom a problem and if the nonresponders can be considered to be a random sample of the population then it can be ignored (see Section 9.2 for a more complete classification of nonresponse). Also, if the units sampled are highly homogeneous then most statisticians would not be too concerned. For example, some laboratory animals have been bred for decades to be quite similar in their genetic background. In contrast, people in most major countries have very different backgrounds and their opinions and genetic makeup can vary greatly.

When only unit nonresponse occurs, the data gathered will look complete in that information is available on all the variables for each case. Suppose in a survey of students 80% of the females respond and 60% of the males respond and the investigator expects males and females to respond differently to a question (X). If in the population 55% are males and 45% are females, then instead of simply getting an overall average of responses for all the students, a weighted average could be reported. For males $w_1 = .55$ and for females $w_2 = .45$. If $\overline{X}_1$ is the mean for males and $\overline{X}_2$ is the mean for females, then a weighted average could be computed as

$$\overline{X} = \frac{\sum w_i \overline{X}_i}{\sum w_i} = \frac{w_1 \overline{X}_1 + w_2 \overline{X}_2}{w_1 + w_2}$$

Another common technique is to assign each observation a weight and the weight is entered into the data set as if it were a variable. Observations are weighted more if they come from subgroups that have a low response rate. This weight may be adjusted so that the sum of the weights equals the sample size. When weighting data, the investigator is assuming that the responders and nonresponders in a subgroup are similar.

In this book, we do not discuss such weighted analyses in detail. A more complete discussion of using weights for adjustment of unit nonresponse can be found in Groves *et al.* (2002) or Little and Rubin (2002). Several types of weights can be used and it is recommended that the reader consider the various options before proceeding. The investigator would need to obtain information on the units in the population to check whether the units in the sample are proportional to the units in the population. For example, in a survey of professionals taken from a listing of society members if the sex, years since graduation, and current employment information is available from both the listing

of the members and the results of the survey, these variables could be used to compute subgroup weights.

The data set should also be screened for item nonresponse. As will be discussed in Section 9.2, most multivariate analyses require complete data on all the variables used in the analysis. If even one variable has a missing value for a case, that case will not be used. Most statistical packages provide programs that indicate how many cases were used in computing common univariate statistics such as means and standard deviations (or report how many cases were missing). Thus it is simple to find which variables have few or numerous missing values.

Some programs can also indicate how many missing values there are for each case. Other programs allow you to transpose or flip your data file so the rows become the columns and the columns become the rows (Table 3.2). Thus the cases and variables are switched as far as the statistical package is concerned. The number of missing values by case can then be found by computing the univariate statistics on the transposed data. Examination of the pattern of missing values is important since it allows the investigator to see if it appears to be distributed randomly or only occurs in some variables. Also, it may have occurred only at the start of the study or close to the end.

Once the pattern of missing data is determined, a decision must be made on how to obtain a complete data set for multivariate analysis. For a first step, most statisticians agree on the following guidelines.

1. If a variable is missing in a very high proportion of cases, then that variable could be deleted.
2. If a case is missing many variables that are crucial to your analysis, then that case could be deleted.

You should also carefully check if there is anything special about the cases that have numerous missing data as this might give you insight into problems in data collection. It might also give some insight into the population to which the results actually apply. Likewise, a variable that is missing in a high proportion of the respondents may be an indication of a special problem. Following the guidelines listed previously can reduce the problems in data analysis but it will not eliminate the problems of reduced efficiency due to discarded data or potential bias due to differences between the data that are complete and the grossly incomplete data. For example, this process may result in a data set that is too small or that is not representative of the total data set. That is, the missing data may not be missing completely at random (see Section 9.2). In such cases, you should consider methods of imputing (or filling-in) the missing data (see Section 9.2 and the books by Rubin, 2004, Little and Rubin, 2002, Schafer, 1997, or Molenberghs and Kenward, 2007).

Item nonresponse can occur in two ways. First, the data may be missing from the start. In this case, the investigator enters a code for missing values

at the time the data are entered into the computer. One option is to enter a symbol that the statistical package being used will automatically recognize as a missing value. For example, a period, an asterisk (*), or a blank space may be recognized as a missing value by some programs. Commonly, a numerical value is used that is outside the range of possible values. For example, for the variable "sex" (with 1 = male and 2 = female) a missing code could be 9. A string of 9s is often used; thus, for the weight of a person 999 could be used as a missing code. Then that value is declared to be missing. For example, for SAS, one could state

if sex = 9, then sex = . ;

Similar statements are used for the other programs. The reader should check the manual for the precise statement.

If the data have been entered into a spreadsheet program, then commonly blanks are used for missing values. In this case, most spreadsheet (and word processor) programs have search-and-replace features that allow you to replace all the blanks with the missing value symbol that your statistical package automatically recognizes. This replacement should be done before the data file is transferred to the statistical package.

The second way in which values can be considered missing is if the data values are beyond the range of the stated maximum or minimum values. For example, if the age of a respondent is entered as 167 and it is not possible to determine the correct value, then the 167 should be replaced with a missing value code so an obviously incorrect value is not used.

Further discussion of the types of missing values and of ways of handling item nonresponse in multivariate data analysis is given in Section 9.2. Here, we will briefly mention one simple method.

The replacement of missing values with the mean value of that variable is a common option in statistical software packages and is the simplest method of imputation. We do not recommend the use of this method when using the multivariate methods given in later chapters. This method results in underestimation of the variances and covariances that are subsequently used in many multivariate analyses.

Detection of outliers

Outliers are observations that appear inconsistent with the remainder of the data set (Barnett and Lewis, 2000). One method for determining outliers has already been discussed, namely, setting minimum and maximum values. By applying these limits, extreme or unreasonable outliers are prevented from entering the data set.

Often, observations are obtained that seem quite high or low but are not impossible. These values are the most difficult ones to cope with. Should they

be removed or not? Statisticians differ in their opinions, from "if in doubt, throw it out" to the point of view that it is unethical to remove an outlier for fear of biasing the results. The investigator may wish to remove an outlier for from the analyses but report them along with the statistical analysis. Another possibility is to run the analyses twice, both with the outliers and without them, to see if they make an appreciable difference in the results. Most investigators would hesitate, for example, to report rejecting a null hypothesis if the removal of an outlier would result in the hypothesis not being rejected.

A review of formal tests for detection of outliers is given in Barnett and Lewis (2000). To make the formal tests you usually must assume normality of the data. Some of the formal tests are known to be quite sensitive to nonnormality and should only be used when you are convinced that this assumption is reasonable. Often an alpha level of 0.10 or 0.15 is used for testing if it is suspected that outliers are not extremely unusual. Smaller values of alpha can be used if outliers are thought to be rare.

The data can be examined one variable at a time by using histograms and box plots if the variable is measured on the interval or ratio scale. A questionable value would be one that is separated from the remaining observations. For nominal or ordinal data, the frequency of each outcome can be noted. If a recorded outcome is impossible, it can be declared missing. If a particular outcome occurs only once or twice, the investigator may wish to consolidate that outcome with a similar one. We will return to the subject of outliers in connection with the statistical analyses starting in Chapter 6, but mainly the discussion in this book is not based on formal tests.

Transformations of the data

Transformations are commonly made either to create new variables with a form more suitable for analysis or to achieve an approximate normal distribution. Here we discuss the first possibility. Transformations to achieve approximate normality are discussed in Chapter 4.

Transformations to create new variables can either be performed as a step in organizing the data or can be included later when the analyses are being performed. It is recommended that they be done as a part of organizing the data. The advantage of this is that the new variables are created once and for all, and sets of instructions for running data analysis from then on do not have to include the data transformation statements. This results in shorter sets of instructions with less repetition and chance for errors when the data are being analyzed. This is almost essential if several investigators are analyzing the same data set.

One common use of transformations occurs in the analysis of questionnaire data. Often the results from several questions are combined to form a new

variable. For example, in studying the effects of smoking on lung function it is common to ask first a question such as:

Have you ever smoked cigarettes? yes___
or no___

If the subjects answer no, they skip a set of questions and go on to another topic. If they answer yes, they are questioned further about the amount in terms of packs per day and length of time they smoked (in years). From this information, a new pack–year variable is created that is the number of years times the average number of packs. For the person who has never smoked, the answer is zero. Transformation statements are used to create the new variable.

Each package offers a slightly different set of transformation statements, but some general options exist. The programs allow you to select cases that meet certain specifications using IF statements. Here for instance, if the response is no to whether the person has ever smoked, the new variable should be set to zero. If the response is yes, then pack–years is computed by multiplying the average amount smoked by the length of time smoked. This sort of arithmetic operation is provided for and the new variable is added to the end of the data set.

Additional options include taking means of a set of variables or the maximum value of a set of variables. Another common arithmetic transformation involves simply changing the numerical values coded for a nominal or ordinal variable. For example, for the depression data set, sex was coded male = 1 and female = 2. In some of the analyses used in this book, we have recoded that to male = 0 and female = 1 by simply subtracting one from the given value.

Saving the results

After the data have been screened for missing values and outliers, and transformations made to form new variables, the results are saved in a master file that can be used for analysis. We recommend that a copy or copies of this master file be made on an external storage device such as a CD or DVD so that it can be stored outside the computer. A summary of decisions made in data screening and transformations used should be stored with the master file. Enough information should be stored so that the investigator can later describe what steps were taken in organizing the data.

If the steps taken in organizing the data were performed by typing in control language, it is recommended that a copy of this control language be stored along with the data sets. Then, should the need arise, the manipulation can be redone by simply editing the control language instructions rather than completely recreating them.

If results are saved interactively (point and click), then it is recommended that multiple copies be saved along the way until you are perfectly satisfied

with the results and that the Windows notepad or some similar memo facility be used to document your steps. Figure 3.1 summarizes the steps taken in data entry and data management.

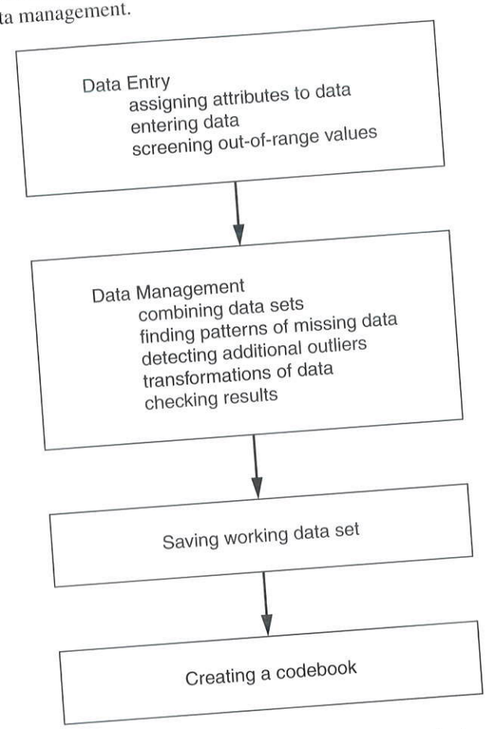

Figure 3.1: *Preparing Data for Statistical Analysis*

3.5 Example: depression study

In this section we discuss a data set that will be used in several succeeding chapters to illustrate multivariate analyses. The depression study itself is described in Chapter 1.

The data given here are from a subset of 294 observations randomly chosen from the original 1000 respondents sampled in Los Angeles. This subset of observations is large enough to provide a good illustration of the statisti-

cal techniques but small enough to be manageable. Only data from the first time period are included. Variables are chosen so that they would be easily understood and would be sensible to use in the multivariate statistical analyses described in Chapters 6–17.

The codebook, the variables used, and the data set are described below.

Codebook

In multivariate analysis, the investigator often works with a data set that has numerous variables, perhaps hundreds of them. An important step in making the data set understandable is to create a written codebook that can be given to all the users. The codebook should contain a description of each variable and the variable name given to each variable for use in the statistical package. Some statistical packages have limits on the length of the variable names so that abbreviations are used. Often blank spaces are not allowed, so dashes or underscores are included. Some statistical packages reserve certain words that may not be used as variable names. The variables should be listed in the same order as they are in the data file. The codebook serves as a guide and record for all users of the data set and for future documentation of the results.

Table 3.3 contains a codebook for the depression data set. In the first column the variable number is listed, since that is often the simplest way to refer to the variables in the computer. A variable name is given next, and this name is used in later data analysis. These names were chosen to be eight characters or less so that they could be used by all the package programs. It is helpful to choose variable names that are easy to remember and are descriptive of the variables, but short to reduce space in the display.

Finally a description of each variable is given in the last column of Table 3.3. For nominal or ordinal data, the numbers used to code each answer are listed. For interval or ratio data, the units used are included. Note that income is given in thousands of dollars per year for the household; thus an income of 15 would be $\$15,000$ per year. Additional information that is sometimes given includes the number of cases that have missing values, how missing values are coded, the largest and smallest value for that variable, simple descriptive statistics such as frequencies for each answer for nominal or ordinal data, and means and standard deviations for interval or ratio data. We note that one package (Stata) can produce a codebook for its users that includes much of the information just described.

Depression variables

The 20 items used in the depression scale are variables 9–28 and are named C1, C2,..., C20. (The wording of each item is given later in the text, in Table 14.2.) Each item was written on a card and the respondent was asked to tell the

Table 3.3: ***Codebook for depression data***

Variable number	Variable name	Description
1	ID	Identification number from 1 to 294
2	SEX	1 = male; 2 = female
3	AGE	Age in years at last birthday
4	MARITAL	1 = never married; 2 = married; 3 = divorced; 4 = separated; 5 = widowed
5	EDUCAT	1 = less than high school; 2 = some high school; 3 = finished high school; 4 = some college; 5 = finished bachelor's degree; 6 = finished master's degree; 7 = finished doctorate
6	EMPLOY	1 = full time; 2 = part time; 3 = unemployed; 4 = retired; 5 = houseperson; 6 = in school; 7 = other
7	INCOME	Thousands of dollars per year
8	RELIG	1 = Protestant; 2 = Catholic; 3 = Jewish; 4 = none; 5 = other
9–28	C1–C20	"Please look at this card and tell me the number that best describes how often you felt or behaved this way during the past week." 20 items from depression scale (already reflected; see text) 0 = rarely or none of the time (less than 1 day); 1 = some or a little of the time (1–2 days); 2 = occasionally or a moderate amount of the time (3–4 days); 3 = most or all of the time (5–7 days)
29	CESD	Sum of C1–20; 0 = lowest level possible; 60 = highest level possible
30	CASES	0 = normal; 1 = depressed, where depressed is CESD$\geq$16
31	DRINK	Regular drinker? 1 = yes; 2 = no
32	HEALTH	General health? 1 = excellent; 2 = good; 3 = fair; 4 = poor
33	REGDOC	Have a regular doctor? 1 = yes; 2 = no
34	TREAT	Has a doctor prescribed or recommended that you take medicine, medical treatments, or change your way of living in such areas as smoking, special diet, exercise, or drinking? 1 = yes; 2 = no
35	BEDDAYS	Spent entire day(s) in bed in last two months? 0 = no; 1 = yes
36	ACUTEILL	Any acute illness in last two months? 0 = no; 1 = yes
37	CHRONILL	Any chronic illness in last year? 0 = no; 1 = yes

interviewer the number that best describes how often he or she felt or behaved this way during the past week. Thus respondents who answered item C2, "I felt depressed," could respond 0–3, depending on whether this particular item applied to them rarely or none of the time (less than 1 day: 0), some or little of the time (1–2 days: 1), occasionally or a moderate amount of the time (3–4 days: 2), or most or all of the time (5–7 days: 3).

Most of the items are worded in a negative fashion, but items C8–C11 are positively worded. For example, C8 is "I felt that I was as good as other people." For positively worded items the scores are **reflected**: that is, a score of 3 is changed to be 0, 2 is changed to 1, 1 is changed to 2, and 0 is changed to 3. In this way, when the total score of all 20 items is obtained by summation of variables C1–C20, a large score indicates a person who is depressed. This sum is the 29th variable, named CESD.

Persons whose CESD score is greater than or equal to 16 are classified as depressed since this value is the common cutoff point used in the literature (Frerichs, Aneshensel and Clark, 1981). These persons are given a score of 1 in variable 30, the CASES variable. The particular depression scale employed here was developed for use in community surveys of noninstitutionalized respondents (Comstock and Helsing, 1976; Radloff, 1977).

Data set

As can be seen by examining the codebook given in Table 3.3 demographic data (variables 2–8), depression data (variables 9–30), and general health data (variables 32–37) are included in this data set. Variable 31, DRINK, was included so that it would be possible to determine if an association exists between drinking and depression. Frerichs *et al.* (1981) have already noted a lack of association between smoking and scores on the depression scale.

The actual data for the first 30 of the 294 respondents are listed in Table 3.4. The rest of the data set, along with the other data sets used in this book, are available on the CRC Press and UCLA web sites (see Appendix A).

3.6 Summary

In this chapter we discussed the steps necessary before statistical analysis can begin. The first of these is the decision of what computer and software packages to use. Once this decision is made, data entry and organizing the data can be started.

Note that investigators often alter the order of these operations. For example, some prefer to check for missing data and outliers and to make transformations prior to combining the data sets. This is particularly true in analyzing longitudinal data when the first data set may be available well before the others. This may also be an iterative process in that finding errors may lead to entering

Table 3.4: *Depression data*

OBS	ID	SEX	AGE	MARITAL	EDUC	EMPLOY	INCOME	RELIG	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	CESD	CASES	DRINK	HEALTH	REGDOC	TREAT	BEDDAYS	ACUTEILL	CHRONILL
1	1	2	68	5	2	4	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	1	1	0	0	1
2	2	1	58	3	4	1	15	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	4	0	1	1	1	1	0	0	1
3	3	2	45	2	3	1	28	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	4	0	1	2	1	1	0	0	0
4	4	2	50	3	3	3	9	1	0	0	0	0	1	1	0	3	0	0	0	0	0	0	0	0	0	0	0	0	5	0	2	1	1	2	0	0	1
5	5	2	33	4	3	1	35	1	0	0	0	0	0	0	0	3	3	0	0	0	0	0	0	0	0	0	0	0	6	0	1	1	1	1	1	1	0
6	6	1	24	2	3	1	11	1	0	0	0	0	0	0	0	0	1	0	0	1	2	0	0	2	1	0	0	0	7	0	1	1	1	1	0	1	1
7	7	2	58	2	2	5	11	1	2	1	1	2	1	0	0	2	2	0	0	0	0	0	3	0	0	0	0	1	15	0	2	3	1	1	0	1	1
8	8	1	22	1	3	1	9	1	0	1	2	0	2	1	0	0	0	0	0	0	0	0	0	1	1	1	1	0	10	0	2	1	2	2	0	1	0
9	9	2	47	2	3	4	23	2	0	1	1	0	0	3	0	0	0	0	0	3	0	3	2	3	0	0	0	0	16	1	1	4	1	1	1	0	1
10	10	1	30	2	2	1	35	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	2	0	0	0
11	11	2	20	1	2	3	25	4	0	0	1	0	1	2	1	0	0	1	0	1	2	2	1	1	2	3	0	0	18	1	1	2	1	2	0	0	0
12	12	2	57	2	3	2	24	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	0	0	4	0	2	2	1	1	1	1	1
13	13	1	39	2	2	1	28	1	1	1	0	0	0	1	0	0	1	0	0	1	0	0	2	0	0	0	0	1	8	0	1	3	1	1	0	1	0
14	14	2	61	5	3	4	13	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	1	4	0	1	1	1	1	1	1	0
15	15	2	23	2	3	1	15	2	0	0	0	0	0	0	1	1	1	2	2	0	0	0	0	0	0	0	0	1	8	0	1	1	1	2	0	0	0
16	16	2	21	1	2	1	6	1	1	1	1	2	0	1	1	1	1	2	2	0	1	1	1	3	1	0	1	0	21	1	1	3	1	1	1	0	1
17	17	2	23	1	4	1	8	1	3	3	3	2	3	3	3	2	2	3	1	2	0	0	2	3	2	0	0	3	42	1	1	1	2	1	1	1	0
18	18	2	55	4	2	3	19	1	1	1	0	1	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	6	0	2	3	1	2	1	1	1
19	19	2	26	1	6	1	15	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	2	2	1	1	0
20	20	1	64	5	2	4	9	4	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	3	0	1	2	1	2	0	0	0
21	21	2	44	1	3	1	6	2	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	3	0	1	1	1	1	0	0	1
22	22	2	25	2	3	1	35	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	4	0	1	2	1	1	0	1	1
23	23	2	72	5	3	4	7	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	2	0	1	2	1	1	0	0	1
24	24	2	61	2	3	1	19	2	0	0	0	0	0	0	0	1	2	0	0	0	1	0	0	0	0	0	0	0	4	0	2	3	1	1	0	0	1
25	25	2	43	3	3	1	6	1	0	0	0	0	0	1	0	0	0	1	0	0	1	1	1	3	2	0	0	0	10	0	2	3	1	1	0	0	0
26	26	2	52	2	2	5	19	2	1	2	1	1	0	1	0	0	3	0	0	0	1	0	0	0	2	0	0	0	12	0	1	3	1	1	0	0	0
27	27	2	23	2	3	5	13	1	0	0	0	0	0	2	2	0	0	0	2	0	0	0	0	0	0	0	0	0	6	0	2	2	1	2	0	1	1
28	28	1	73	4	2	4	5	2	0	1	0	2	0	1	0	2	1	1	1	0	0	0	0	0	0	0	0	0	9	0	1	3	1	1	0	0	0
29	29	2	34	2	3	2	19	2	0	2	2	2	0	1	2	2	1	1	1	2	0	2	3	3	2	0	0	2	28	1	1	2	1	2	0	0	0
30	30	2	34	2	3	1	20	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	2	1	2	0	0	1

new data to replace erroneous values. Again, we stress saving the results on CDs or DVDs or some other external storage device after each set of changes.

Six statistical packages — R, S-PLUS, SAS, SPSS, Stata, and STATISTICA — were noted as the packages used in this book. In evaluating a package it is often helpful to examine the data entry and data manipulation features they offer. The tasks performed in data entry and organization are often much more difficult and time consuming than running the statistical analyses, so a package that is easy and intuitive to use for these operations is a real help. If the package available to you lacks needed features, then you may wish to perform these operations in one of the spreadsheet or relational database packages and then transfer the results to your statistical package.

3.7 Problems

3.1 Enter the data set given in Table 8.1, Chemical companies' financial performance (Section 8.3), using a data entry program of your choice. Make a codebook for this data set.

3.2 Using the data set entered in Problem 3.1 delete the P/E variable for the Dow Chemical company and D/E for Stauffer Chemical and Nalco Chemical in a way appropriate for the statistical package you are using. Then, use the missing value features in your statistical package to find the missing values and replace them with an imputed value.

3.3 Using your statistical package, compute a scatter diagram of income versus employment status from the depression data set. From the data in this table, decide if there are any adults whose income is unusual considering their employment status. Are there any adults in the data set whom you think are unusual?

3.4 Transfer a data set from a spreadsheet program into your statistical software package.

3.5 Describe the person in the depression data set who has the highest total CESD score.

3.6 Using a statistical software program, compute histograms for mothers' and fathers' heights and weights from the lung function data set described in Section 1.2 and in Section A.5 of Appendix A. The data and codebook can be obtained from the web site listed in Section A.7 of Appendix A. Describe cases that you consider to be outliers.

3.7 From the lung function data set (Problem 3.6), determine how many families have one child, two children, and three children between the ages of 7 and 18.

3.8 For the lung function data set, produce a two-way table of gender of child 1 versus gender of child 2 (for families with at least two children). Comment.