© Andreas Gursky, VG BILD-KUNST, Bonn

FIFTH EDITION

THE BASICS OF SOCIAL RESEARCH

EARL BABBIE

Ryan Mcvay/The Image Bank/Getty Images

14 Quantitative Data Analysis

What You'll Learn in This Chapter

Often, social data are converted to numerical form for statistical analyses. In this chapter, we'll begin with the process of quantifying data, then turn to analysis. Quantitative analysis may be descriptive or explanatory; it may involve one, two, or several variables. We begin our examination of quantitative analyses with some simple but powerful ways of manipulating data in order to attain research conclusions.

In this chapter . . .

What do you think?

Earl Babbie

In Chapter 13, we saw several inherent shortcomings in quantitative data. These shortcomings centered primarily on standardization and superficiality in the face of a social reality that is varied and deep. Can anything meaningful be learned from data that sacrifice meaningful detail in order to permit numerical manipulations?

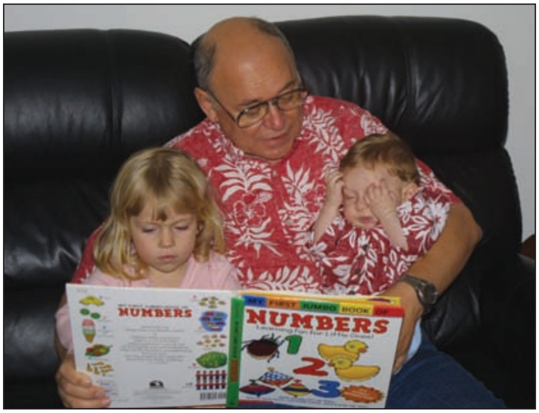
See the "What do you think? Revisited" box toward the end of the chapter.

INTRODUCTION

In Chapter 13, we saw some of the logic and techniques by which social researchers analyze the qualitative data they've collected. This chapter will examine **quantitative analysis**, or the techniques by which researchers convert data to a numerical form and subject it to statistical analyses.

To begin, we'll look at *quantification*—the process of converting data to a numerical format. This involves converting social science data into a *machine-readable form*—a form that can be read and manipulated by computers and similar machines used in quantitative analysis.

The rest of the chapter will present the logic and some of the techniques of quantitative data analysis—starting with the simplest case, univariate analysis, which involves one variable, then discussing bivariate analysis, which involves two variables. We'll move on to a brief introduction to multivariate analysis, or the examination of several variables simultaneously, such as *age,*



Aaron Babbie

Some students take to statistics more readily than do others.

quantitative analysis The numerical representation and manipulation of observations for the purpose of describing and explaining the phenomena that those observations reflect.

education, and *prejudice,* and then we'll move to a discussion of sociological diagnostics. Finally, we'll look at the ethics of quantitative data analysis.

Before we can do any sort of analysis, we need to quantify our data. Let's turn now to the basic steps involved in converting data into machine-readable forms amenable to computer processing and analysis.

QUANTIFICATION OF DATA

Today, quantitative analysis is almost always done by computer programs such as SPSS and MicroCase. For those programs to work their magic, they must be able to read the data you've collected in your research. If you've conducted a survey, for example, some of your data are inherently numerical: age or income, for instance. Whereas the writing and check marks on a questionnaire are qualitative in nature, a scribbled age is easily converted to quantitative data.

Other data are also easily quantified: Transforming male and female into "1" and "2" is hardly rocket science. Researchers can also easily assign numerical representations to such variables as *religious affiliation, political party,* and *region of the country.*

Some data are more challenging, however. If a survey respondent tells you that he or she thinks the biggest problem facing Woodbury, Vermont, is "the disintegrating ozone layer," the computer can't process that response numerically. You must translate by coding the responses. We've already discussed coding in connection with content analysis (Chapter 11) and again in connection with qualitative data analysis (Chapter 13). Now we look at coding specifically for quantitative analysis, which differs from the other two primarily in its goal of converting raw data into numbers.

As with content analysis, the task of quantitative coding is to reduce a wide variety of idiosyncratic items of information to a more limited set of attributes composing a variable. Suppose, for example, that a survey researcher asks respondents, "What is your occupation?" The responses to such a question will vary considerably. Although it will be possible to assign each reported occupation a separate numerical code, this procedure will not facilitate analysis, which typically depends on several subjects having the same attribute.

The variable *occupation* has many preestablished coding schemes. One such scheme distinguishes professional and managerial occupations, clerical occupations, semiskilled occupations, and so forth. Another scheme distinguishes different sectors of the economy: manufacturing, health, education, commerce, and so forth. Still others combine both of these schemes. Using an established coding scheme gives you the advantage of being able to compare your research results with those of other studies.

To learn more about preestablished coding schemes, visit the Bureau of Labor Statistics to learn about their Standard Occupational Classification: stats.bls.gov/soc/soc_majo.htm.

The occupational coding scheme you choose should be appropriate for the theoretical concepts being examined in your study. For some studies, coding all occupations as either white-collar or blue-collar might suffice. For others, self-employed and not self-employed might do. Or a peace researcher might wish to know only whether the occupation depended on the defense establishment or not.

Although the coding scheme should be tailored to meet particular requirements of the analysis, you should keep one general guideline in mind. If the data are coded to maintain a great deal of detail, code categories can always be combined during an analysis that does not require such detail. If the data are coded into relatively few, gross categories, however, you'll have no way during analysis to recreate the original detail. To keep your options open, it's a good idea to code your data in greater detail than you plan to use in the analysis.

Developing Code Categories

There are two basic approaches to the coding process. First, you may begin with a relatively well-developed coding scheme, derived from your research purpose. Thus, as suggested previously, the peace researcher might code occupations in terms of their relationship to the defense establishment. You might also use an existing coding scheme so that you can compare your findings with those of previous research.

The alternative method is to generate codes from your data, as discussed in Chapter 13. Let's say we've asked students in a self-administered campus survey to say what they believe is the biggest problem facing their college today. Here are a few of the answers they might have written in.

Tuition is too high
Not enough parking spaces
Faculty don't know what they are doing
Advisors are never available
Not enough classes offered
Cockroaches in the dorms
Too many requirements
Cafeteria food is infected
Books cost too much
Not enough financial aid

Take a minute to review these responses and see whether you can identify some categories represented. Realize that there is no right answer; several coding schemes might be generated from these answers.

Let's start with the first response: "Tuition is too high." What general areas of concern does that response reflect? One obvious possibility is "Financial Concerns." Are there other responses that would fit into that category? Table 14-1 shows which of the questionnaire responses could fit.

In more general terms, the first answer can also be seen as reflecting nonacademic concerns. This categorization would be relevant if your research interest included the distinction between academic and nonacademic concerns. If that were the case, the responses might be coded as shown in Table 14-2.

TABLE 14-1 Student Responses That Can Be Coded "Financial Concerns"

	Financial Concerns
Tuition is too high	X
Not enough parking spaces	
Faculty don't know what they are doing	
Advisors are never available	
Not enough classes offered	
Cockroaches in the dorms	
Too many requirements	
Cafeteria food is infected	
Books cost too much	X
Not enough financial aid	X

Notice that I didn't code the response "Books cost too much" in Table 14-2, because this concern could be seen as representing both of the categories. Books are part of the academic program, but their cost is not. This signals the need to refine the coding scheme we're developing. Depending on our research purpose, we might be especially interested in identifying any problems that had an academic element; hence we'd code this one

TABLE 14-2 Student Concerns Coded as "Academic" and "Nonacademic"

	Academic	Nonacademic
Tuition is too high		X
Not enough parking spaces		X
Faculty don't know what they are doing	X	
Advisors are never available	X	
Not enough classes offered	X	
Cockroaches in the dorms		X
Too many requirements	X	
Cafeteria food is infected		X
Books cost too much		
Not enough financial aid		X

"Academic." Just as reasonably, however, we might be more interested in identifying nonacademic problems and would code the response accordingly. Or, as another alternative, we might create a separate category for responses that involved both academic and nonacademic matters.

As yet another alternative, we might want to separate nonacademic concerns into those involving administrative matters and those dealing with campus facilities. Table 14-3 shows how the first ten responses would be coded in that event.

As these few examples illustrate, there are many possible schemes for coding a set of data. Your choices should match your research purposes and reflect the logic that emerges from the data themselves. Often, you'll find yourself modifying the code categories as the coding process proceeds. Whenever you change the list of categories, however, you must review the data already coded to see whether changes are in order.

TABLE 14-3 Nonacademic Concerns Coded as "Administrative" or "Facilities"

	Academic	Administrative	Facilities
Tuition is too high		X	
Not enough parking spaces			X
Faculty don't know what they are doing	X		
Advisors are never available	X		
Not enough classes offered	X		
Cockroaches in the dorms			X
Too many requirements	X		
Cafeteria food is infected			X
Books cost too much	X		
Not enough financial aid		X	

Like the set of attributes composing a variable, and like the response categories in a closed-ended questionnaire item, code categories should be both exhaustive and mutually exclusive. Every piece of information being coded should fit into one and only one category. Problems arise whenever a given response appears to fit equally into more than one code category or whenever it fits into no category: Both signal a mismatch between your data and your coding scheme.

If you're fortunate enough to have assistance in the coding process, you'll need to train your coders in the definitions of code categories and show them how to use those categories properly. To do so, explain the meaning of the code categories and give several examples of each. To make sure your coders fully understand what you have in mind, code several cases ahead of time. Then ask your coders to code the same cases without knowing how you coded them. Finally, compare your coders' work with your own. Any discrepancies will indicate an imperfect communication of your coding scheme to your coders. Even with perfect agreement between you and your coders, however, it's best to check the coding of at least a portion of the cases throughout the coding process.

If you're not fortunate enough to have assistance in coding, you should still obtain some verification of your own reliability as a coder. Nobody's perfect, especially a researcher hot on the trail of a finding. Suppose that you're studying an emerging cult and that you have the impression that people who do not have a regular family will be the most likely to regard the new cult as a family substitute. The danger is that whenever you discover a subject who reports no family, you'll unconsciously try to find some evidence in the subject's comments that the cult is a substitute for family. If at all possible, then, get someone else to code some of your cases to see whether that person makes the same assignments you made.

Codebook Construction

The end product of the coding process in quantitative analysis is the conversion of data items into numerical codes. These codes represent attributes composing variables, which, in turn, are assigned locations within a data file. A **codebook** is a document that describes the locations of variables and lists the assignments of codes to the attributes composing those variables.

A codebook serves two essential functions. First, it is the primary guide used in the coding process. Second, it is your guide for locating variables and interpreting codes in your data file during analysis. If you decide to correlate two variables as a part of your analysis of your data, the codebook tells you where to find the variables and what the codes represent.

Figure 14-1 is a partial codebook created from two variables from the General Social Survey. Though there is no one right format for a codebook, this example presents some of the common elements.

Notice first that each variable is identified by an abbreviated variable name: POLVIEWS, ATTEND. We can determine the religious service attendance of respondents, for example, by referencing ATTEND. This example uses the format established by the General Social Survey, which has been carried over into SPSS. Other data sets and/or analysis programs might format variables differently. Some use numerical codes in place of abbreviated names, for example. You must, however, have some identifier that will allow you to locate and use the variable in question.

Next, every codebook should contain the full definition of the variable. In the case of a questionnaire, the definition consists of the exact wordings of the questions asked, because, as we've seen, the wording of questions strongly influences the answers returned. In the case of POLVIEWS, you know that respondents were given the several political categories and asked to pick the one that best fit them.

The codebook also indicates the attributes composing each variable. In POLVIEWS, for example, the political categories just mentioned serve as these attributes: "Extremely liberal," "Liberal," "Slightly liberal," and so forth.

Finally, notice that each attribute also has a numeric label. Thus, in POLVIEWS, "Extremely liberal" is code category 1. These numeric codes are used in various manipulations of the data. For example, you might decide to combine categories 1 through 3 (all the "liberal" responses). It's easier to do this with code numbers than with lengthy names.

Data Entry

In addition to transforming data into quantitative form, researchers interested in quantitative analysis also need to convert data into a machine-readable format, so that computers can read and manipulate the data. There are many ways of accomplishing this step, depending on the original form of your data and also the computer program you'll use for analyzing the data. I'll simply introduce you to the process here. If you find yourself undertaking this task, you should be able to tailor your work to the particular data source and program you're using.

If your data have been collected by questionnaire, you might do your coding on the questionnaire itself. Then, data-entry specialists (including yourself) could enter the data into, say, an SPSS data matrix or into an Excel spreadsheet that would later be imported into SPSS.

Sometimes, social researchers use optical scan sheets for data collection. These sheets can be fed into machines that will convert the black marks into data, which can be imported into the analysis program. This procedure only works with subjects who are comfortable using such sheets, and it's usually limited to closed-ended questions.

Sometimes, data entry occurs in the process of data collection. In computer-assisted telephone interviewing (CATI), for example, the

codebook The document used in data processing and analysis that tells the location of different data items in a data file. Typically, the codebook identifies the locations of data items and the meaning of the codes used to represent different attributes of variables.

POLVIEWS

We hear a lot of talk these days about liberals and conservatives. I'm going to show you a seven-point scale on which the political views that people might hold are arranged from extremely liberal—point 1—to extremely conservative—point 7. Where would you place yourself on this scale?

1. Extremely liberal
2. Liberal
3. Slightly liberal
4. Moderate, middle of the road
5. Slightly conservative
6. Conservative
7. Extremely conservative
8. Don't know
9. No answer

ATTEND

How often do you attend religious services?

0. Never
1. Less then once a year
2. About once or twice a year
3. Several times a year
4. About once a month
5. 2–3 times a month
6. Nearly every week
7. Every week
8. Several times a week
9. Don't know, No answer

FIGURE 14-1 **Partial Codebook.**

interviewer keys responses directly into the computer, where the data are compiled for analysis (see Chapter 9). Even more effortlessly, online surveys can be constructed so that the respondents enter their own answers directly into the accumulating database, without the need for an intervening interviewer or data-entry person.

Once data have been fully quantified and entered into the computer, researchers can begin quantitative analysis. Let's look at the three cases mentioned at the start of this chapter: univariate, bivariate, and multivariate analyses.

UNIVARIATE ANALYSIS

The simplest form of quantitative analysis, **univariate analysis**, involves describing a case in terms of a single variable—specifically, the distribution of attributes that compose it. For example, if *sex* were measured, we would look at how many of the subjects were men and how many were women.

Distributions

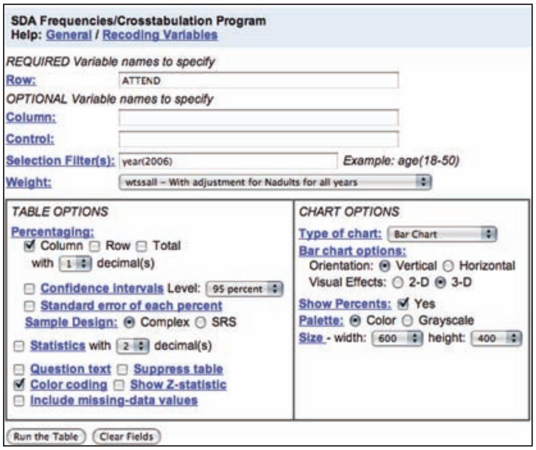
The most basic format for presenting univariate data is to report all individual cases, that is, to list the attribute for each case under study in terms of the variable in question. Let's take as an example the General Social Survey (GSS) data on attendance at religious services, ATTEND.

Figure 14-2 shows how you could request these data, using the Berkeley SDA online analysis program introduced earlier in the book. You can access this program at sda.berkeley.edu/cgi-bin32/hsda?harcsda+gss06.

In the figure you'll see that ATTEND has been entered as the Row variable, and I have specified a Selection Filter to limit the analysis to the data collected in the 2006 GSS. Notice, also, that I've selected Bar Chart as the Type of Chart, have asked for 3-D effects and have asked to see the percentages. The consequence of this will be apparent shortly.

Table 14-4 represents the tabular response to our request. We see, for example, that 1,009 of

univariate analysis The analysis of a single variable, for purposes of description. Frequency distributions, averages, and measures of dispersion are examples of univariate analysis, as distinguished from *bivariate* and *multivariate analysis.*



SDA Frequencies/Crosstabulation Program
Help: General / Recoding Variables

REQUIRED Variable names to specify

Row: ATTEND

OPTIONAL Variable names to specify

Column:

Control:

Selection Filter(s): year(2006) *Example: age(18-50)*

Weight: wtssall – With adjustment for Nadults for all years

TABLE OPTIONS

Percentaging:
☑ Column □ Row □ Total
with 1 decimal(s)

□ Confidence intervals Level: 95 percent
□ Standard error of each percent
Sample Design: ● Complex ○ SRS

□ Statistics with 2 decimal(s)

□ Question text □ Suppress table
☑ Color coding □ Show Z-statistic
□ Include missing-data values

CHART OPTIONS

Type of chart: Bar Chart
Bar chart options:
Orientation: ● Vertical ○ Horizontal
Visual Effects: ○ 2-D ● 3-D
Show Percents: ☑ Yes
Palette: ● Color ○ Grayscale
Size - width: 600 height: 400

FIGURE 14-2 Requesting a Univariate Analysis of ATTEND.

the 4,492 respondents, or 22.5 percent, say they never attend worship services. As we move down the table, we see that 19 percent say they attend every week. To simplify the results, we might want to combine the last three categories and say that 31.1 percent attend "About weekly."

A description of the number of times that the various attributes of a variable are observed in a sample is called a **frequency distribution**. Sometimes it's easiest to see a frequency distribution in a graph. Figure 14-3 was created by SDA based on the specifications in the chart options section of Figure 14-2. The vertical scale on the left side of the graph indicates the percentages selecting each of the answers that are displayed along the horizontal axis of the graph. Take a minute to notice how the percentages in Table 14-4 correspond to the heights of the bars in Figure 14-3.

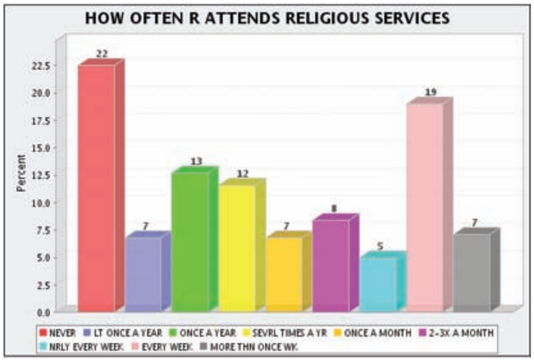


FIGURE 14-3 Bar Chart of GSS ATTEND, 2006.

This program also offers other graphical possibilities. In Figure 14-2, you could have specified "Pie Chart" instead of "Bar Chart" as the type of chart desired. Figure 14-4 shows the way the data would have been presented in that case.

Central Tendency

Beyond simply reporting the overall distribution of values, sometimes called the marginal frequencies or just the marginals, you may choose to present your data in the form of an **average**, or measure of central tendency. You're already familiar with the concept of central tendency from the many kinds of averages you use in everyday life to express the "typical" value of a variable. For instance, in baseball a batting average of .300 says that a batter gets a hit three out of every ten opportunities on average. Over the course of a season, a hitter might go through extended periods without getting any hits at all and go through other periods when he or she gets a bunch of hits all at once. Over time, though, the central tendency of the batter's performance can be expressed as getting three hits in every ten chances. Similarly, your grade point average expresses the "typical" value of all your grades taken together, even though some of them might

frequency distribution A description of the number of times the various attributes of a variable are observed in a sample. The report that 53 percent of a sample were men and 47 percent were women would be a simple example of a frequency distribution.

average An ambiguous term generally suggesting typical or normal—a central tendency. The *mean*, *median*, and *mode* are specific examples of mathematical averages.

TABLE 14-4 Attendance at Worship Services, 2006

ATTEND	How Often R Attends Religious Services		
Value Label	Value	Frequency	Percent
NEVER	0	1,009	22.5
LT ONCE A YEAR	1	305	6.8
ONCE A YEAR	2	571	12.7
SEVRL TIMES A YR	3	522	11.6
ONCE A MONTH	4	307	6.8
2–3X A MONTH	5	378	8.4
NRLY EVERY WEEK	6	224	5.0
EVERY WEEK	7	856	19.0
MORE THN ONCE WK	8	321	7.1
	Total	4,492	100.0

be A's, others B's, and one or two might be C's (I know you never get anything lower than a C).

Averages like these are more properly called the *arithmetic mean* (the result of dividing the sum of the values by the total number of cases). The **mean** is only one way to measure central tendency or "typical" values. Two other options are the **mode** (the most frequently occurring attribute) and the **median** (the middle attribute in the ranked distribution of observed attributes). Here's how the three averages would be calculated from a set of data.

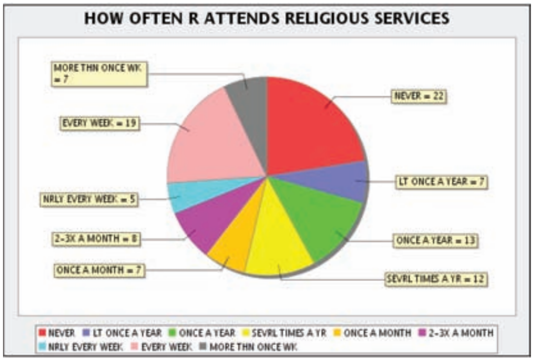


FIGURE 14-4 **Pie Chart of GSS ATTEND, 2006.**

Suppose you're conducting an experiment that involves teenagers as subjects. They range in age from 13 to 19, as indicated in the following table:

Age	Number
13	3
14	4
15	6
16	8
17	4
18	3
19	3

mean An average computed by summing the values of several observations and dividing by the number of observations. If you now have a grade point average of 4.0 based on 10 courses, and you get an F in this course, your new grade point (mean) average will be 3.6.

mode An average representing the most frequently observed value or attribute. If a sample contains 1,000 Protestants, 275 Catholics, and 33 Jews, "Protestant" is the modal category.

median An average representing the value of the "middle" case in a rank-ordered set of observations. If the ages of five men are 16, 17, 20, 54, and 88, the median would be 20. (The mean would be 39.)

Now that you've seen the actual ages of the 31 subjects, how old would you say they are in general, or "on average"? Let's look at three different ways you might answer that question.

The easiest average to calculate is the mode, the most frequent value. As you can see, there were more 16-year-olds (eight of them) than any other age, so the modal age is 16, as indicated in Figure 14-5. Technically, the modal age is the category "16," which may include some people who are closer to 17 than 16 but who haven't yet reached that birthday.

Figure 14-5 also demonstrates the calculation of the mean. There are three steps: (1) multiply each age by the number of subjects who have that age, (2) total the results of all those multiplications, and (3) divide that total by the number of subjects.

In the case of age, a special adjustment is needed. As indicated in the discussion of the mode, those who call themselves "13" actually range from exactly 13 years old to those just short of 14. It is reasonable to assume, moreover, that as a group the "13-year-olds" in the country are evenly distributed within that one-year span, making their average age 13.5 years. This is true for each of the age groups. Hence, it's appropriate to add 0.5 years to the final calculation, making the mean age 16.37, as indicated in Figure 14-5.

The third measure of central tendency, the median, represents the "middle" value: Half are above it, half below. If we had the precise ages of each subject (for example, 17 years and 124 days), we'd be able to arrange all 31 subjects in order by age, and the median for the whole group would be the age of the middle subject.

As you can see, however, we do not know precise ages; our data constitute "grouped data" in this regard. For example, three people who are not precisely the same age have been grouped in the category "13-year-olds."

Figure 14-5 illustrates the logic of calculating a median for grouped data. Because there are 31 subjects altogether, the "middle" subject would be subject number 16 if they were arranged by age—15 teenagers would be younger and 15 older. Look at the bottom portion of Figure 14-5, and you'll see that the middle person is one of the eight 16-year-olds. In the enlarged view of that group, we see that number 16 is the third from the left.

Because we do not know the precise ages of the subjects in this group, the statistical convention here is to assume they are evenly spread along the width of the group. In this instance, the possible ages of the subjects go from 16 years and no days to 16 years and 364 days. Strictly speaking, the range, then, is 364/365 days. As a practical matter, it's sufficient to call it one year.

If the eight subjects in this group were evenly spread from one limit to the other, they would be one-eighth of a year apart from each other—a 0.125-year interval. Look at the illustration and you'll see that if we place the first subject half the interval from the lower limit and add a full interval to the age of each successive subject, the final one is half an interval from the upper limit.

What we've done is calculate, hypothetically, the precise ages of the eight subjects, assuming their ages were spread out evenly. Having done this, we merely note the age of the middle subject—16.31—and that is the median age for the group.

Whenever the total number of subjects is an even number, of course, there is no middle case. To get the median, you merely calculate the mean of the two values on either side of the midpoint in the ranked data. Suppose, for example, that there was one more 19-year-old in our sample, giving us a total of 32 cases. The midpoint would then fall between subjects 16 and 17. The median would therefore be calculated as (16.31 + 16.44)/2 = 16.38.

As you can see in Figure 14-5, the three measures of central tendency produce three different values for this set of data, which is often (but not necessarily) the case. Which measure, then, best represents the "typical" value? More generally, which measure of central tendency should you prefer? The answer depends on the nature of your data and the purpose of your analysis. For example, whenever means are presented, you should be aware that they are susceptible to extreme values—a few very large or very small numbers. As only one example, the (mean) average person in Redmond, Washington, has a net

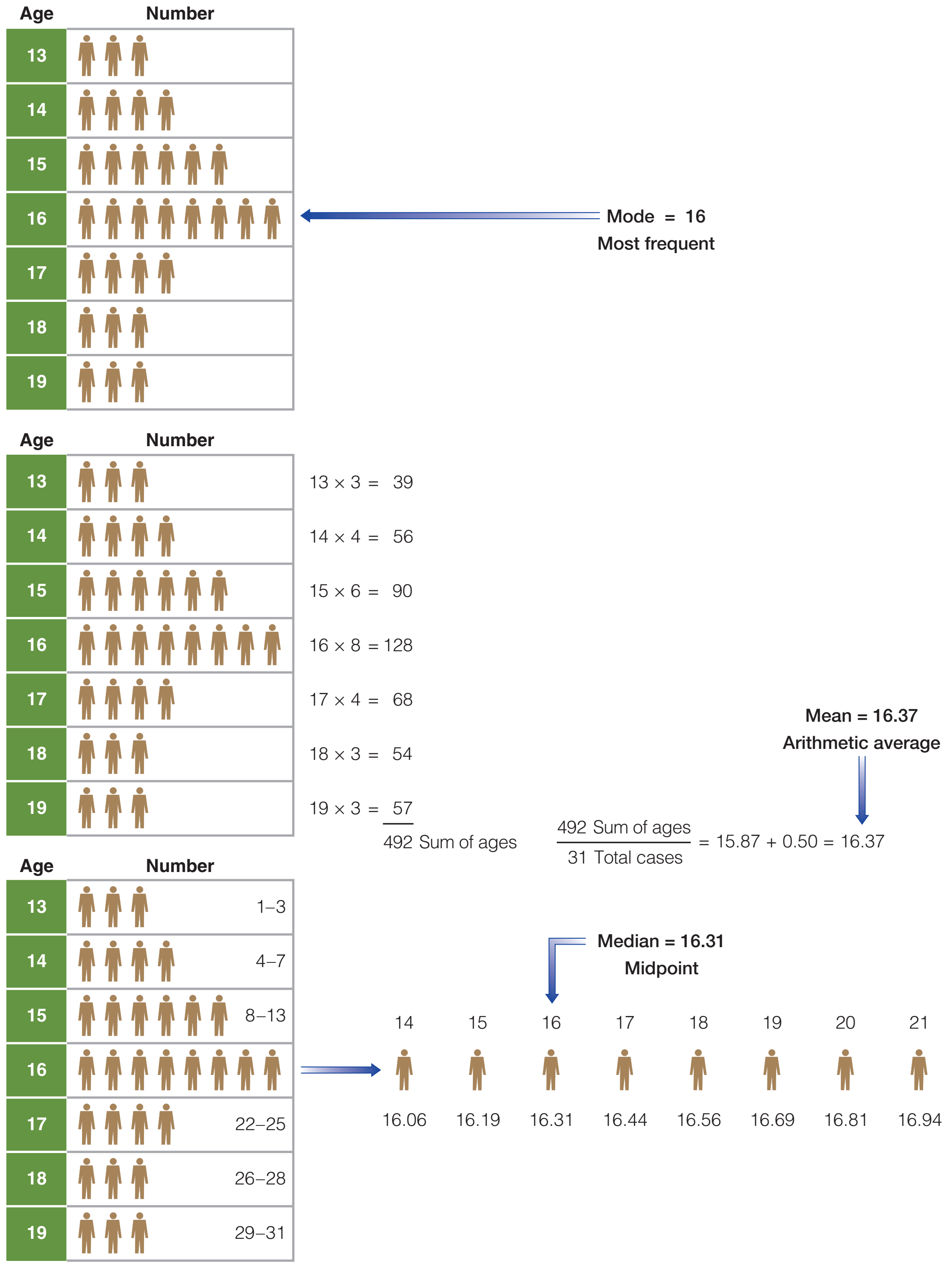


FIGURE 14-5 Three "Averages."

worth in excess of a million dollars. If you were to visit Redmond, however, you would not find that the "average" resident lives up to your idea of a millionaire. The very high mean reflects the influence of one extreme case among Redmond's 40,000 residents—Bill Gates of Microsoft, who has a net worth (at the time this is being written) of tens of billions of dollars. Clearly, the median wealth would give you a more accurate picture of the residents of Redmond as a whole.

This example should illustrate the need to choose carefully among the various measures of central tendency. A course or textbook in statistics will give you a fuller understanding of the variety of situations in which each is appropriate.

Dispersion

Averages offer readers the advantage of reducing the raw data to the most manageable form: A single number (or attribute) can represent all the detailed data collected in regard to the variable. This advantage comes at a cost, of course, because the reader cannot reconstruct the original data from an average. Summaries of the dispersion of responses can somewhat alleviate this disadvantage.

Dispersion refers to the way values are distributed around some central value, such as an average. The simplest measure of dispersion is the range: the distance separating the highest from the lowest value. Thus, besides reporting that our subjects have a mean age of 15.87, we might also indicate that their ages range from 13 to 19.

A more sophisticated measure of dispersion is the **standard deviation**. This measure was briefly mentioned in Chapter 7 as the standard error of a sampling distribution. Essentially, the standard deviation is an index of the amount of variability in a set of data. A higher standard deviation means that the data are more dispersed; a lower standard deviation means that they are more bunched together. Figure 14-6 illustrates the basic idea. Notice that the professional golfer not only has a lower mean score but is also more consistent—represented by the lower standard deviation. The duffer, on the other hand, has a higher average and is also less consistent: sometimes doing much better, sometimes much worse.

There are many other measures of dispersion. In reporting intelligence test scores, for example, researchers might determine the interquartile range, the range of scores for the middle 50 percent of subjects. If the top one-fourth had scores ranging from 120 to 150, and if the bottom one-fourth had scores ranging from 60 to 90, the report might say that the interquartile range was from 90 to 120 (or 30 points) with a mean score of, let's say, 102.

Continuous and Discrete Variables

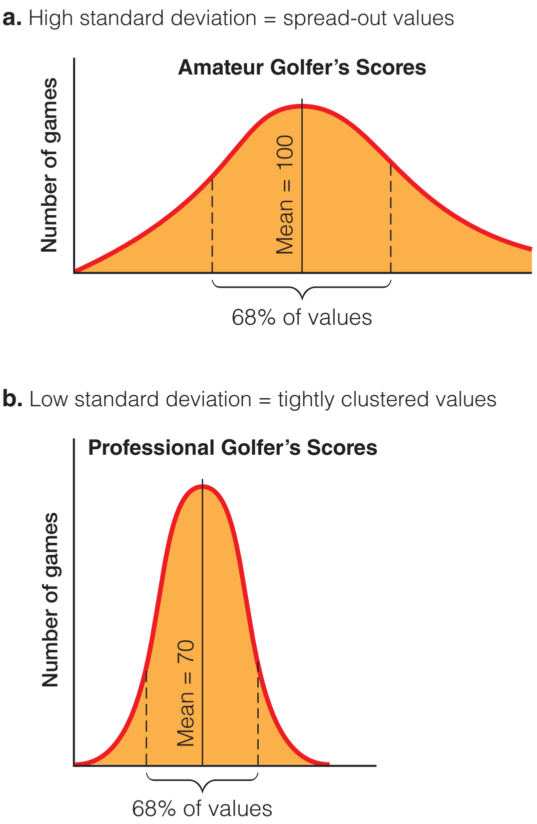
The preceding calculations are not appropriate for all variables. To understand this point, we must distinguish between two types of variables: continuous and discrete. A **continuous variable** (or ratio variable) increases steadily in tiny fractions. An example is *age*, which increases steadily with each increment of time. A **discrete variable** jumps from category to category without intervening steps. Examples include *sex*, *military rank*, or *year in college* (you go from being a sophomore to a junior in one step).

dispersion The distribution of values around some central value, such as an average. The range is a simple example of a measure of dispersion. Thus, we may report that the mean age of a group is 37.9, and the range is from 12 to 89.

standard deviation A measure of dispersion around the mean, calculated so that approximately 68 percent of the cases will lie within plus or minus one standard deviation from the mean, 95 percent will lie within plus or minus two standard deviations, and 99.9 percent will lie within three standard deviations. Thus, for example, if the mean age in a group is 30 and the standard deviation is 10, then 68 percent have ages between 20 and 40. The smaller the standard deviation, the more tightly the values are clustered around the mean; if the standard deviation is high, the values are widely spread out.

continuous variable A variable whose attributes form a steady progression, such as *age* or *income*. Thus, the ages of a group of people might include 21, 22, 23, 24, and so forth and could even be broken down into fractions of years.

discrete variable A variable whose attributes are separate from one another, or discontinuous, as in the case of *sex* or *religious affiliation*. In other words, there is no progression from male to female in the case of *sex*.


a. High standard deviation = spread-out values

b. Low standard deviation = tightly clustered values

FIGURE 14-6 **High and Low Standard Deviations.**

In analyzing a discrete variable—a nominal or ordinal variable, for example—some of the techniques discussed previously do not apply. Strictly speaking, modes should be calculated for nominal data, medians for interval data, and means for ratio data, not for nominal data (see Chapter 5). If the variable in question is *sex*, for example, raw numbers (23 of the cross-dressing outlaw bikers in our sample are women) or percentages (7 percent are women) can be appropriate and useful analyses, but neither a median nor a mean would make any sense. Calculating the mode would be legitimate, though not very revealing, because it would only tell us "most were men." However, the mode for data on *religious affiliation* might be more interesting, as in "most people in the United States are Protestant."

Detail versus Manageability

In presenting univariate and other data, you'll be constrained by two goals. On the one hand, you should attempt to provide your reader with the fullest degree of detail regarding those data. On the other hand, the data should be presented in a manageable form. As these two goals often directly conflict, you'll find yourself continually seeking the best compromise between them. One useful solution is to report a given set of data in more than one form. In the case of age, for example, you might report the distribution of ungrouped ages plus the mean age and standard deviation.

As you can see from this introductory discussion of univariate analysis, this seemingly simple matter can be rather complex. In any event, the lessons of this section pave the way for a consideration of subgroup comparisons and bivariate analyses.

SUBGROUP COMPARISONS

Univariate analyses describe the units of analysis of a study and, if they are a sample drawn from some larger population, allow us to make descriptive inferences about the larger population. Bivariate and multivariate analyses are aimed primarily at explanation. Before turning to explanation, however, we should consider the case of subgroup description.

Often it's appropriate to describe subsets of cases, subjects, or respondents. Here's a simple example from the General Social Survey. In 2006, respondents were asked, "Should marijuana be made legal?" In response, 34.9 percent said it should and 65.1 percent said it shouldn't. Table 14-5 presents the responses given to this question by respondents in different age categories.

Notice that the subgroup comparisons tell us how different groups in the population responded to this question. You can undoubtedly see a pattern in the results, though possibly not exactly what you expected; we'll return to that in a moment. First, let's see how another set of subgroups answered this question.

TABLE 14-5 Marijuana Legalization by Age of Respondents, 2006

	Under 21	21–35	36–54	55 and older
Should be legalized	34%	37%	38%	29%
Should not be legalized	66	63	62	71
100% =	(57)	(574)	(704)	(513)

Source: General Social Survey, 2006, National Opinion Research Center.

TABLE 14-6 Marijuana Legalization by Political Orientation, 2006

	Should Legalize	Should Not Legalize	100% =
Extremely liberal	50%	50	(59)
Liberal	52%	48	(197)
Slightly liberal	48%	52	(217)
Moderate	36%	64	(669)
Slightly conservative	34%	66	(292)
Conservative	17%	83	(294)
Extremely conservative	17%	83	(73)

Source: General Social Survey, 2006, National Opinion Research Center.

Table 14-6 presents attitudes toward legalizing marijuana by different political subgroups, based on whether respondents characterized themselves as conservative or liberal. Before looking at the table, you might try your hand at hypothesizing what the results are likely to be and why. Notice that I've changed the direction of percentaging this table, to make it easier to read. To compare the subgroups in this case, you would read down the columns, not across them.

Before examining the logic of causal analysis, let's consider another example of subgroup comparisons—one that will let us address some table-formatting issues.

"Collapsing" Response Categories

"Textbook examples" of tables are often simpler than you'll typically find in published research reports or in your own analyses of data, so this section and the next one address two common problems and suggest solutions.

Let's begin by turning to Table 14-7, which reports data collected in a multinational poll conducted by the *New York Times,* CBS News, and the *Herald Tribune* in 1985, concerning attitudes about the United Nations. The question reported in Table 14-7 deals with general attitudes about the way the UN was handling its job.

Here's the question: How do people in the five nations reported in Table 14-7 compare in their support for the kind of job the UN was doing? As you review the table, you may find there are simply so many numbers that it's hard to see any meaningful pattern.

Part of the problem with Table 14-7 lies in the relatively small percentages of respondents selecting the two extreme response categories: the UN is doing a very good or a very poor job. Furthermore, although it might be tempting to read only the second line of the table (those saying "good job"), that would be improper. Looking at only the second row, we would conclude that West Germany and the United States were the most positive (46 percent) about the UN's performance, followed closely by France (45 percent), with Britain (39 percent) less positive than any of those three and Japan (11 percent) the least positive of all.

This procedure is inappropriate in that it ignores all those respondents who gave the most positive answer of all: "very good job." In a situation like this, you should combine or "collapse" the two ends of the range of variation. In this instance, combine "very good" with "good" and "very poor" with "poor." If you were to do this in the analysis of your own data, it would be wise to add the raw frequencies together and recompute percentages for the combined categories, but in analyzing a published table such as this one, you can simply add the percentages, as illustrated by the results shown in Table 14-8.

With the collapsed categories illustrated in Table 14-8, we can now rather easily read across

TABLE 14-7 Attitudes toward the United Nations: "How is the UN doing in solving the problems it has had to face?"

	West Germany	Britain	France	Japan	United States
Very good job	2%	7%	2%	1%	5%
Good job	46	39	45	11	46
Poor job	21	28	22	43	27
Very poor job	6	9	3	5	13
Don't know	26	17	28	41	10

Source: "5-Nation Survey Finds Hope for U.N.," *New York Times*, June 26, 1985, p. 6.

the several national percentages of people who said the UN was doing at least a good job. Now the United States appears the most positive; Germany, Britain, and France are only slightly less positive and are nearly indistinguishable from one another; and Japan stands alone in its quite low assessment of the UN's performance. Although the conclusions to be drawn now do not differ radically from what we might have concluded from simply reading the second line of Table 14-7, we should note that Britain now appears relatively more supportive.

Here's the risk I'd like to spare you. Suppose you had hastily read the second row of Table 14-7 and noted that the British had a somewhat lower assessment of the job the UN was doing than was true of people in the United States, West Germany, and France. You might feel obliged to think up an explanation for why that was so—possibly creating an ingenious psychohistorical theory about the painful decline of the once powerful and dignified British Empire. Then, once you had touted your "theory" about, someone else might point out that a proper reading of the data would show the British were actually not really less positive than the other three nations. This is not a hypothetical risk. Errors like these happen frequently, but they can be avoided by collapsing answer categories where appropriate.

Handling "Don't Knows"

Tables 14-7 and 14-8 illustrate another common problem in the analysis of survey data. It's usually a good idea to give people the option of saying "don't know" or "no opinion" when asking for their opinions on issues. But what do you do with those answers when you analyze the data?

Notice there is a good deal of variation in the national percentages saying "don't know" in this instance, ranging from only 10 percent in the United States to 41 percent in Japan. The presence of substantial percentages saying they don't know can confuse the results of tables like these. For example, were the Japanese so much less likely to say the UN was doing a good job simply because so many didn't express any opinion?

Here's an easy way to recalculate percentages, with the "don't knows" excluded. Look at the first column of percentages in Table 14-8: West Germany's answers to the question about the UN's performance. Notice that 26 percent of the

TABLE 14-8 Collapsing Extreme Categories

	West Germany	Britain	France	Japan	United States
Good job or better	48%	46%	47%	12%	51%
Poor job or worse	27	37	25	48	40
Don't know	26	17	28	41	10

respondents said they didn't know. This means that those who said "good" or "bad" job—taken together—represent only 74 percent (100 minus 26) of the whole. If we divide the 48 percent saying "good job or better" by 0.74 (the proportion giving any opinion), we can say that 65 percent "of those with an opinion" said the UN was doing a good or very good job (48%/0.74 = 65%).

Table 14-9 presents the whole table with the "don't knows" excluded. Notice that these new data offer a somewhat different interpretation than do the previous tables. Specifically, it would now appear that France and West Germany were the most positive in their assessments of the UN, with the United States and Britain a bit lower. Although Japan still stands out as lowest in this regard, it has moved from 12 percent to 20 percent positive.

At this point, having seen three versions of the data, you may be asking yourself, Which is the right one? The answer depends on your purpose in analyzing and interpreting the data. For example, if it is not essential for you to distinguish "very good" from "good," it makes sense to combine them, because it's easier to read the table.

Whether to include or exclude the "don't knows" is harder to decide in the abstract. It may be a very important finding that such a large percentage of the Japanese had no opinion—if you wanted to find out whether people were familiar with the work of the UN, for example. On the other hand, if you wanted to know how people might vote on an issue, it might be more appropriate to exclude the "don't knows" on the assumption that they wouldn't vote or that ultimately they would be likely to divide their votes between the two sides of the issue.

In any event, the truth contained within your data is that a certain percentage said they didn't know and the remainder divided their opinions in whatever manner they did. Often, it's appropriate to report your data in both forms—with and without the "don't knows"—so your readers can also draw their own conclusions. Of course, you yourself will be a reader of such tables, drawn up by others, and knowing the logic behind constructing them will help you be a savvy consumer of quantitative data.

Numerical Descriptions in Qualitative Research

Although this chapter deals primarily with quantitative research, the discussions are also relevant to qualitative studies. Numerical testing can often verify the findings of in-depth, qualitative studies. Thus, for example, when David Silverman wanted to compare the cancer treatments received by patients in private clinics with those in Britain's National Health Service, he primarily chose in-depth analyses of the interactions between doctors and patients:

> My method of analysis was largely qualitative and . . . I used extracts of what doctors and patients had said as well as offering a brief ethnography of the setting and of certain behavioural data. In addition, however, I constructed a coding form which enabled me to collate a number of crude measures of doctor and patient interactions. (1993:163)

Not only did the numerical data fine-tune Silverman's impressions based on his qualitative observations, but his in-depth understanding of the situation allowed him to craft an ever-more appropriate quantitative analysis. Listen to the interaction between qualitative and quantitative approaches in this lengthy discussion:

TABLE 14-9 Omitting the "Don't Knows"

	West Germany	Britain	France	Japan	United States
Good job or better	65%	55%	65%	20%	57%
Poor job or worse	35%	45%	35%	81%	44%

> My overall impression was that private consultations lasted considerably longer than those held in the NHS clinics. When examined, the data indeed did show that the former were almost twice as long as the latter (20 minutes as against 11 minutes) and that the difference was statistically highly significant. However, I recalled that, for special reasons, one of the NHS clinics had abnormally short consultations. I felt a fairer comparison of consultations in the two sectors should exclude this clinic and should only compare consultations taken by a single doctor in both sectors. This subsample of cases revealed that the difference in length between NHS and private consultations was now reduced to an average of under 3 minutes. This was still statistically significant, although the significance was reduced. Finally, however, if I compared only new patients seen by the same doctor, NHS patients got 4 minutes more on the average—34 minutes as against 30 minutes in the private clinic. (1993:163–64)

This example further demonstrates the special power that can be gained from a combination of approaches in social research. The combination of qualitative and quantitative analyses can be especially potent.

BIVARIATE ANALYSIS

In contrast to univariate analysis, subgroup comparisons involve two variables. In this respect, subgroup comparisons constitute a kind of **bivariate analysis**—that is, an analysis of two variables simultaneously. However, as with univariate analysis, the purpose of subgroup comparisons is largely descriptive. Most bivariate analysis in social research adds another element: determining relationships between the variables themselves. Thus, univariate analysis and subgroup comparisons focus on describing the people (or other units of analysis) under study, whereas bivariate analysis focuses on the variables and their empirical relationships.

Table 14-10 could be regarded as an instance of subgroup comparison: It independently describes the attendance of men and women at religious services, as reported in the 2006 General Social Survey. It shows—comparatively and descriptively—that the women under study attended religious services more often than did the men. However, the same table, seen as an explanatory bivariate analysis, tells a somewhat different story. It suggests that the variable *sex* has an effect on the variable *religious service attendance.* That is, we can view the behavior as a dependent variable that is partially determined by the independent variable, *sex.*

Explanatory bivariate analyses, then, involve the "variable language" introduced in Chapter 1. In a subtle shift of focus, we are no longer talking about men and women as different subgroups but about *sex* as a variable: one that has an influence on other variables. The theoretical interpretation of Table 14-10 might be taken from Charles Glock's Comfort Hypothesis as discussed in Chapter 2:

1. Women are still treated as second-class citizens in U.S. society.
2. People denied status gratification in the secular society may turn to religion as an alternative source of status.
3. Hence, women should be more religious than men.

The data presented in Table 14-10 confirm this reasoning. Thirty-five percent of the women attend religious services weekly, as compared with 26 percent of the men.

Adding the logic of causal relationships among variables has an important implication for the construction and reading of percentage tables. One of the chief bugaboos for new-data analysts is deciding on the appropriate "direction of percentaging" for any given table. In Table 14-10, for example, I've divided the group of subjects into two subgroups—men and women—and then

bivariate analysis The analysis of two variables simultaneously, for the purpose of determining the empirical relationship between them. The construction of a simple percentage table or the computation of a simple correlation coefficient are examples of bivariate analyses.

TABLE 14-10 Religious Attendance Reported by Men and Women, 2006

	Men	Women
Weekly	26%	35%
Less often	74	65
100% =	(2,049)	(2,443)

Source: General Social Survey, 2006, National Opinion Research Center.

described the behavior of each subgroup. That is the correct method for constructing this table. Notice, however, that we could—however inappropriately—construct the table differently. We could first divide the subjects into different degrees of religious attendance and then describe each of those subgroups in terms of the percentage of men and women in each. This method would make no sense in terms of explanation, however. Table 14-10 suggests that your sex will affect your frequency of religious service attendance. Had we used the other method of construction, the table would suggest that your religious service attendance affects whether you are a man or a woman—which makes no sense. Your behavior cannot determine your sex.

A related problem complicates the lives of new-data analysts. How do you read a percentage table? There is a temptation to read Table 14-10 as follows: "Of the women, only 35 percent attended religious services weekly, and 65 percent said they attended less often; therefore, being a woman makes you less likely to attend religious services frequently." This is, of course, an incorrect reading of the table. Any conclusion that sex—as a variable—has an effect on religious service attendance must hinge on a comparison between men and women. Specifically, we compare the 35 percent with the 26 percent and note that women are more likely than men to attend religious services weekly. The comparison of subgroups, then, is essential in reading an explanatory bivariate table.

In constructing and presenting Table 14-10, I've used a convention called *percentage down.* This term means that you can add the percentages down each column to total 100 percent. You read this form of table across a row. For the row labeled "Weekly," what percentage of the men attend weekly? What percentage of the women attend weekly?

The direction of percentaging in tables is arbitrary, and some researchers prefer to percentage across, as I did in Table 14-6. They would organize Table 14-10 so that "Men" and "Women" were shown on the left side of the table, identifying the two rows, and "Weekly" and "Less often" would appear at the top to identify the columns. The actual numbers in the table would be moved around accordingly, and each row of percentages would total 100 percent. In that case, you would read the table down a column, still asking what percentage of men and women attended frequently. The logic and the conclusion would be the same in either case; only the form would differ.

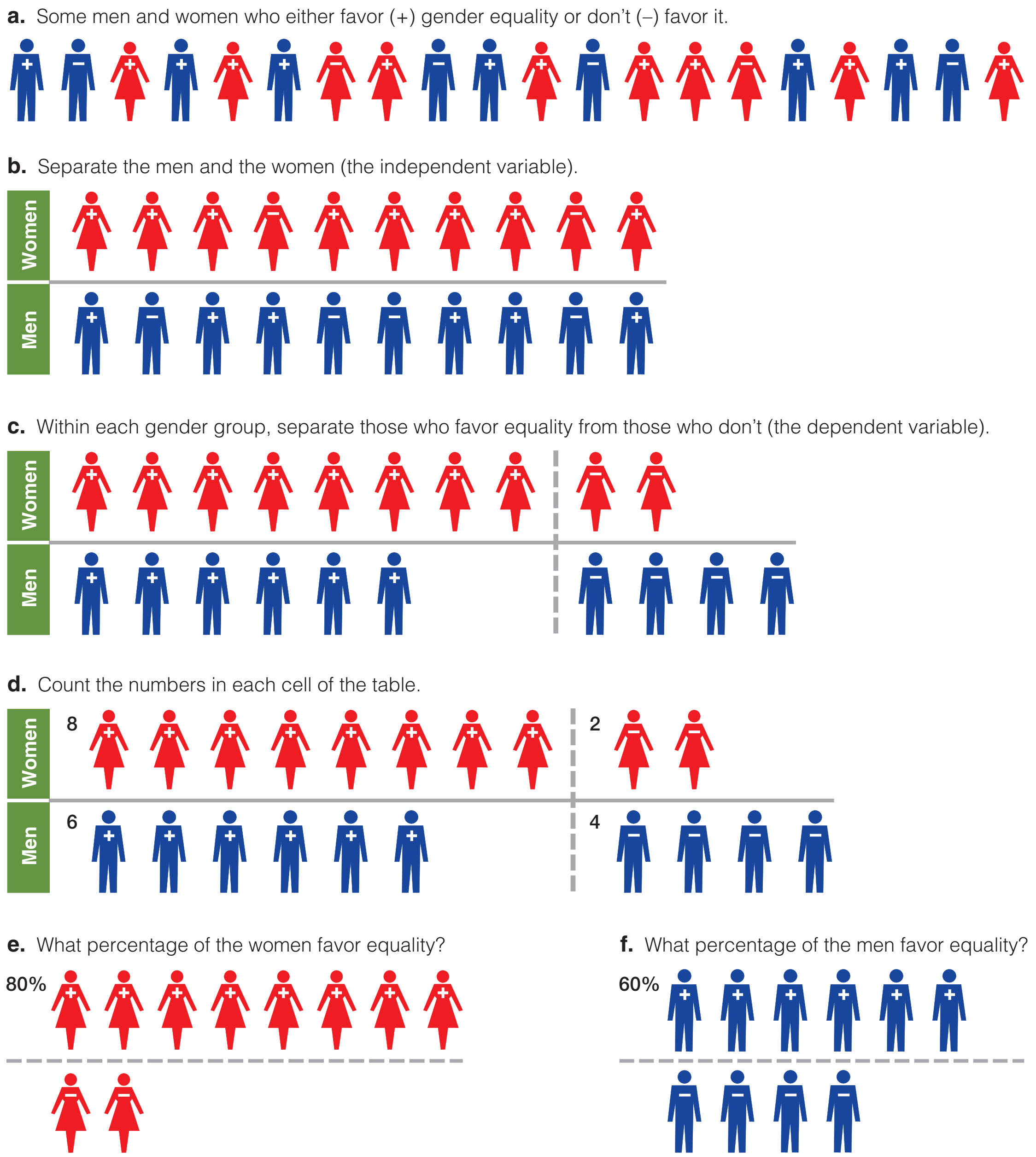
In reading a table that someone else has constructed, therefore, you need to find out in which direction it has been percentaged. Usually this will be labeled or be clear from the logic of the variables being analyzed. As a last resort, however, you should add the percentages in each column and each row. If each of the columns totals 100 percent, the table has been percentaged down. If the rows total 100 percent each, it has been percentaged across. The rule, then, is as follows:

1. If the table is percentaged down, read across.
2. If the table is percentaged across, read down.

Percentaging a Table

Figure 14-7 reviews the logic by which we create percentage tables from two variables. I've used as variables *sex* and *attitude toward equality for men and women.*

Here's another example. Suppose we're interested in learning something about newspaper editorial policies regarding the legalization of marijuana. We undertake a content analysis of editorials on this subject that have appeared during a given year in a sample of daily newspapers across the nation. Each editorial has



a. Some men and women who either favor (+) gender equality or don't (–) favor it.

b. Separate the men and the women (the independent variable).

c. Within each gender group, separate those who favor equality from those who don't (the dependent variable).

d. Count the numbers in each cell of the table.

e. What percentage of the women favor equality?

f. What percentage of the men favor equality?

g. Conclusions.

While a majority of both men and women favored gender equality, women were more likely than men to do so.

Thus, gender appears to be one of the causes of attitudes toward sexual equality.

	Women	Men
Favor equality	80%	60%
Don't favor equality	20	40
Total	**100%**	**100%**

FIGURE 14-7 **Percentaging a Table.**

been classified as favorable, neutral, or unfavorable toward the legalization of marijuana. Perhaps we wish to examine the relationship between editorial policies and the types of communities in which the newspapers are published, thinking that rural newspapers might be more conservative in this regard than urban ones. Thus, each newspaper (hence, each editorial) has been classified in terms of the population of the community in which it is published.

Table 14-11 presents hypothetical data describing the editorial policies of rural and urban newspapers. Note that the unit of analysis in this example is the individual editorial. Table 14-11 tells us that there were 127 editorials about marijuana in our sample of newspapers published in communities with populations under 100,000. (Note that this cutting point is chosen for simplicity of illustration and does not mean that *rural* refers to a community of less than 100,000 in any absolute sense.) Of these, 11 percent (14 editorials divided by a base of 127) were favorable toward legalization of marijuana, 29 percent were neutral, and 60 percent were unfavorable. Of the 438 editorials that appeared in our sample of newspapers published in communities of more than 100,000 residents, 32 percent (140 editorials) were favorable toward legalizing marijuana, 40 percent were neutral, and 28 percent were unfavorable.

When we compare the editorial policies of rural and urban newspapers in our imaginary study, we find—as expected—that rural newspapers are less favorable toward the legalization of marijuana than are urban newspapers. We determine this by noting that a larger percentage (32 percent) of the urban editorials were favorable than the percentage of rural ones (11 percent). We might note as well that more rural than urban editorials were unfavorable (60 percent compared with 28 percent). Note that this table assumes that the size of a community might affect its newspapers' editorial policies on this issue, rather than that editorial policy might affect the size of communities.

TABLE 14-11 Hypothetical Data Regarding Newspaper Editorials on the Legalization of Marijuana

Editorial Policy toward Legalizing Marijuana	Community Size	
	Under 100,000	Over 100,000
Favorable	11%	32%
Neutral	29	40
Unfavorable	60	28
100% =	(127)	(438)

Constructing and Reading Bivariate Tables

Let's now review the steps involved in the construction of explanatory bivariate tables:

1. The cases are divided into groups according to the attributes of the independent variable.
2. Each of these subgroups is then described in terms of attributes of the dependent variable.
3. Finally, the table is read by comparing the independent variable subgroups with each other in terms of a given attribute of the dependent variable.

Let's repeat the analysis of sex and attitude on gender equality following these steps. For the reasons outlined previously, *sex* is the independent variable; *attitude toward gender equality* constitutes the dependent variable. Thus, we proceed as follows:

1. The cases are divided into men and women.
2. Each sex subgrouping is described in terms of approval or disapproval of gender equality.
3. Men and women are compared in terms of the percentages approving of gender equality.

In the example of editorial policies regarding the legalization of marijuana, *size of community* is the independent variable, and a *newspaper's editorial policy* the dependent variable. The table would be constructed as follows:

1. Divide the editorials into subgroups according to the sizes of the communities in which the newspapers are published.
2. Describe each subgroup of editorials in terms of the percentages favorable, neutral, or unfavorable toward the legalization of marijuana.
3. Compare the two subgroups in terms of the percentages favorable toward the legalization of marijuana.

Bivariate analyses typically have an explanatory causal purpose. These two hypothetical examples have hinted at the nature of causation as social scientists use it.

Tables such as the ones we've been examining are commonly called **contingency tables**: Values of the dependent variable are contingent on (depend on) values of the independent variable. Although contingency tables are common in social science, their format has never been standardized. As a result, you'll find a variety of formats in research literature. As long as a table is easy to read and interpret, there's probably no reason to strive for standardization. However, there are several guidelines that you should follow in the presentation of most tabular data:

1. A table should have a heading or a title that succinctly describes what is contained in the table.
2. The original content of the variables should be clearly presented—in the table itself if at all possible or in the text with a paraphrase in the table. This information is especially critical when a variable is derived from responses to an attitudinal question, because the meaning of the responses will depend largely on the wording of the question.
3. The attributes of each variable should be clearly indicated. Though complex categories will have to be abbreviated, their meaning should be clear in the table and, of course, the full description should be reported in the text.
4. When percentages are reported in the table, the base on which they are computed should be indicated. It's redundant to present all the raw numbers for each category, because these could be reconstructed from the percentages and the bases. Moreover, the presentation of both numbers and percentages often confuses a table and makes it more difficult to read.
5. If any cases are omitted from the table because of missing data ("no answer," for example), their numbers should be indicated in the table.

Although I have introduced the logic of causal, bivariate analysis in terms of percentage tables, many other formats are appropriate for this topic. Scatterplot graphs are one possibility, providing a visual display of the relationship between two variables. For an engaging example of this, you might check out the GapMinder software available on the web. Using countries as the unit of analysis, you can examine the relationship between birthrate and infant mortality, for example. In fact, you can watch the relationship develop over time.

You can find GapMinder at tools.google.com/gapminder/.

INTRODUCTION TO MULTIVARIATE ANALYSIS

The logic of **multivariate analysis**, or the analysis of more than two variables simultaneously, can be seen as an extension of bivariate analysis. Specifically, we can construct multivariate tables on the basis of a more complicated subgroup description by following essentially the same steps outlined for bivariate tables. Instead

contingency table A format for presenting the relationships among variables as percentage distributions; typically used to reveal the effects of the independent variable on the dependent variable.

multivariate analysis The analysis of the simultaneous relationships among several variables. Examining simultaneously the effects of *age*, *sex*, and *social class* on *religiosity* would be an example of multivariate analysis.

of one independent variable and one dependent variable, however, we'll have more than one independent variable. Instead of explaining the dependent variable on the basis of a single independent variable, we'll seek an explanation through the use of more than one independent variable.

Let's return to the example of religious attendance. Suppose we believe that age would also affect such behavior (Glock's Comfort Hypothesis suggests that older people are more religious than younger people). As the first step in table construction, we would divide the total sample into subgroups based on the attributes of both independent variables simultaneously: younger men, older men, younger women, and older women. Then the several subgroups would be described in terms of the dependent variable, *religious service attendance,* and comparisons would be made. Table 14-12, from an analysis of the 2006 General Social Survey data, is the result.

Table 14-12 has been percentaged down and therefore should be read across. The interpretation of this table warrants several conclusions:

1. Among both men and women, older people attend religious services more often than do younger people. Among women, 27 percent of those under 40 and 41 percent of those 40 and older attend religious services weekly. Among men, the respective figures are 19 and 31 percent.
2. Within each age group, women attend slightly more frequently than men. Among those respondents under 40, 27 percent of the women attend weekly, compared with 19 percent of the men. Among those 40 and over, 41 percent of the women and 31 percent of the men attend weekly.
3. As measured in the table, age appears to have a greater effect on attendance at religious services than does sex.
4. Age and sex have independent effects on religious service attendance. Within a given attribute of one independent variable, different attributes of the second still affect behaviors.
5. Similarly, the two independent variables have a cumulative effect on behaviors. Older women attend the most often (41 percent), and younger men attend the least often (19 percent).

TABLE 14-12 Multivariate Relationship: Religious Service Attendance, Sex, and Age, 2006

	"How often do you attend religious services?"			
	Under 40		40 and Older	
	Men	Women	Men	Women
About weekly*	19%	27%	31%	41%
Less often	81	73	69	59
100% =	(832)	(958)	(1,211)	(1,477)

*About weekly = "More than once a week," "Weekly," and "Nearly every week."

Source: General Social Survey, 2006, National Opinion Research Center.

Before I conclude this section, it will be useful to note an alternative format for presenting such data. Several of the tables presented in this chapter are somewhat inefficient. When the dependent variable, *religious attendance,* is dichotomous (having exactly two attributes), knowing one attribute permits the reader to reconstruct the other easily. Thus, if we know that 27 percent of the women under 40 attend religious services weekly, then we know automatically that 73 percent attend less often. So, reporting the percentages who attend less often is unnecessary.

On the basis of this recognition, Table 14-12 could be presented in the alternative format of Table 14-13. In Table 14-13, the percentages of people saying they attend religious services about weekly are reported in the cells representing the intersections of the two independent variables. The numbers presented in parentheses below each percentage represent the number of cases on which the percentages are based. Thus, for example, the reader knows there are 958 women under 40 years of age in the sample, and 27 percent of them attend religious services weekly. We can calculate from this that 259 of those 958 women attend weekly and that the other 699 younger women (or 73 percent) attend less frequently. This new table is easier to read than the former one, and it does not sacrifice any detail.

TABLE 14-13 A Simplification of Table 14-12

	Percent Who Attend about Weekly	
	Men	Women
Under 40	19 (832)	27 (958)
40 and Older	31 (1,211)	41 (1,477)

Source: General Social Survey, 2006, National Opinion Research Center.

SOCIOLOGICAL DIAGNOSTICS

The multivariate techniques we are now exploring can serve as powerful tools for diagnosing social problems. They can be used to replace opinions with facts and to settle ideological debates with data analysis.

For an example, let's return to the issue of sex and income. Many explanations have been advanced to account for the long-standing pattern of women in the labor force earning less than men. One explanation is that, because of traditional family patterns, women as a group have participated less in the labor force and many only begin working outside the home after completing certain child-rearing tasks. Thus, women as a group will probably have less seniority at work than will men, and income increases with seniority. An important 1984 study by the Census Bureau showed this reasoning to be partly true, as Table 14-14 shows.

Table 14-14 indicates, first of all, that job tenure did indeed affect income. Among both men and women, those with more years on the job earned more. This is seen by reading down the first two columns of the table.

TABLE 14-14 Sex, Job Tenure, and Income, 1984 (Full-time workers 21–64 years of age)

	Average Hourly Income		
Years Working with Current Employer	Men	Women	Women/Men Ratio
Less than 2 years	$8.46	$6.03	0.71
2 to 4 years	$9.38	$6.78	0.72
5 to 9 years	$10.42	$7.56	0.73
10 years more	$12.38	$7.91	0.64

Source: U.S. Bureau of the Census, Current Population Reports, Series P-70, No. 10, *Male-Female Differences in Work Experience, Occupation, and Earning, 1984* (Washington, DC: U.S. Government Printing Office, 1987), 4.

The table also indicates that women earned less than men, regardless of job seniority. This can be seen by comparing average wages across the rows of the table, and the ratio of women-to-men wages is shown in the third column. Thus, years on the job was an important determinant of earnings, but seniority did not adequately explain the pattern of women earning less than men. In fact, we see that women with 10 or more years on the job earned substantially less ($7.91/hour) than men with less than two years ($8.46/hour).

Although years on the job did not fully explain the difference between men's and women's pay, there are other possible explanations: level of education, child care responsibilities, and so forth. The researchers who calculated Table 14-14 also examined some of the other variables that might reasonably explain the differences in pay without representing gender discrimination, including these:

- Number of years in the current occupation
- Total years of work experience (any occupation)
- Whether they have usually worked full time
- Marital status
- Size of city or town they live in
- Whether covered by a union contract
- Type of occupation
- Number of employees in the firm

- Whether private or public employer
- Whether they left previous job involuntarily
- Time spent between current and previous job
- Race
- Whether they have a disability
- Health status
- Age of children
- Whether they took an academic curriculum in high school
- Number of math, science, and foreign language classes in high school
- Whether they attended private or public high school
- Educational level achieved
- Percentage of women in the occupation
- College major

Each of the variables listed here might reasonably affect earnings and, if women and men differ in these regards, could help to account for male/female income differences. When all these variables were taken into account, the researchers were able to account for 60 percent of the discrepancy between the incomes of men and women. The remaining 40 percent, then, is a function of other "reasonable" variables and/or prejudice. This kind of conclusion can be reached only by examining the effects of several variables at the same time—that is, through multivariate analysis.

I hope this example shows how the logic implicit in day-to-day conversations can be represented and tested in a quantitative data analysis like this. See "Keeping Humanity in Focus" for more on gender discrimination in the workplace.

KEEPING HUMANITY IN FOCUS

Transsexuals are those individuals who choose permanently and biologically change their sex though surgery and hormones. Clearly such a radical life change brings many interesting adjustments and challenges. Many aspects of the transition would make for interesting studies, and Kristen Schilt has taken a particularly intriguing tack.

While many kinds of research point to the disadvantaged status of women in the workplace, Schilt has used the experiences of transsexuals, in this case female-to-male, to personalize the impact of gender. In many cases, the subjects changed their sex while maintaining the same job in their employing organization. Following their sex change, these men tended to receive pay raises and increased authority. In other studies, male-to-female transsexuals reported just the opposite experience. Personal accounts such as these flesh out statistical studies that consistently show women earning less than men, even when they do the same work.

Source: Kristen Schilt, "Just One of the Guys? How Transmen Make Gender Visible in the Workplace," *Gender and Society* 20, no. 4 (2006): 465–90.

As another example of multivariate data analysis in real life, consider the common observation that minority group members are more likely to be denied bank loans than are white applicants. A counterexplanation might be that the minority applicants in question were more likely to have had a prior bankruptcy or that they had less collateral to guarantee the requested loan—both reasonable bases for granting or denying loans. However, the kind of multivariate analysis we've just examined could easily resolve the disagreement.

Let's say we look only at those who have not had a prior bankruptcy and who have a certain level of collateral. Are whites and minorities equally likely to get the requested loan? We could conduct the same analysis in subgroups determined by level of collateral. If whites and minorities were equally likely to get their loans in each of the subgroups, we would need to conclude that there was no ethnic discrimination. If minorities were still less likely to get their loans, however, that would indicate that bankruptcy and collateral differences were not the explanation—strengthening the case that discrimination was at work.

All this should make clear that social research can play a powerful role in serving in the human community. It can help us determine the current state of affairs and can often point the way to where we want to go.

Welcome to the world of sociological diagnostics!

ETHICS AND QUANTITATIVE DATA ANALYSIS

In Chapter 13, I pointed out that the subjectivity present in qualitative data analysis increases the risk of biased analyses, which experienced researchers learn to avoid. Some think, however, that quantitative analyses are not susceptible to subjective biases. Unfortunately, this isn't so. Even the most mathematically explicit analysis yields ample room for defining and measuring variables in ways that encourage one finding over another, and quantitative analysts need to guard against this. Sometimes, the careful specification of hypotheses in advance can offer protection, although this can also inhibit a full exploration of what data can tell us.

The quantitative analyst has an obligation to report any formal hypotheses and other expectations that didn't pan out. Suppose that you think that a particular variable will prove to be a powerful cause of gender prejudice, but your data analysis contradicts that expectation. You should report the lack of correlation, because such information is useful to others who conduct research on this topic. Although it would be more satisfying to discover what causes prejudice, it's quite important to know what doesn't cause it.

The protection of subjects' privacy is as important in quantitative analysis as in qualitative analysis. However, with quantitative methods it's often easier to collect and record data in ways that make subject identification more difficult. However, the first time public officials demand that you reveal the names of student subjects who reported using illegal drugs in a survey, this issue will take on more salience. (Don't reveal the names, by the way. If necessary, burn the questionnaires—"accidentally.")

What do you think?
REVISITED

This chapter began with a question about whether anything meaningful or useful could be learned from the analysis of data that have been stripped of many details in order to permit statistical manipulation. The answer, we've seen, is an unqualified "yes."

Quantitative analysis can be a tool for social change. For instance, calculating the average incomes of men and women or of whites and minorities can demonstrate the inequalities that exist for people doing exactly the same job. Such quantitative analyses can overpower anecdotal evidence about particular women or minorities who earn large salaries. We've also seen that quantitative analyses of qualitative phenomena, such as voting intentions, can be done with precision and utility.

The key lesson is that both qualitative and quantitative research are legitimate and powerful approaches to understanding social life. They are particularly useful, moreover, when used together.

Main Points

Introduction

- Most data are initially qualitative: They must be quantified to permit statistical analysis.
- Quantitative analysis involves the techniques by which researchers convert data to a numerical form and subject it to statistical analyses.

Quantification of Data

- Some data, such as age and income, are intrinsically numerical.
- Often, quantification involves coding into categories that are then given numerical representations.
- Researchers may use existing coding schemes, such as the Census Bureau's categorization of occupations, or develop their own coding categories. In either case, the coding scheme must be appropriate for the nature and objectives of the study.
- A codebook is the document that describes the identifiers assigned to different variables and the codes assigned to represent the attributes of those variables.

Univariate Analysis

- Univariate analysis is the analysis of a single variable. Because univariate analysis does not involve the relationships between two or more variables, its purpose is descriptive rather than explanatory.
- Several techniques allow researchers to summarize their original data to make them more manageable while maintaining as much of the original detail as possible. Frequency distributions, averages, grouped data, and measures of dispersion are all ways of summarizing data concerning a single variable.

Subgroup Comparisons

- Subgroup comparisons can be used to describe similarities and differences among subgroups with respect to some variable.
- Collapsing response categories and handling "don't knows" are two techniques for presenting and interpreting data.

Bivariate Analysis

- Bivariate analysis focuses on relationships between variables rather than comparisons of groups. Bivariate analysis explores the statistical association between the independent variable and the dependent variable. Its purpose is usually explanatory rather than merely descriptive.
- The results of bivariate analyses often are presented in the form of contingency tables, which are constructed to reveal the effects of the independent variable on the dependent variable.

Introduction to Multivariate Analysis

- Multivariate analysis is a method of analyzing the simultaneous relationships among several variables. It may also be used to understand the relationship between two variables more fully.
- The logic and techniques of quantitative research can be valuable to qualitative researchers.

Sociological Diagnostics

- Sociological diagnostics is a quantitative analysis technique for determining the nature of social problems such as ethnic or gender discrimination.

Ethics and Quantitative Data Analysis

- Unbiased analysis and reporting is as much an ethical concern in quantitative analysis as in qualitative analysis.

- Subjects' privacy must be protected in quantitative data analysis and reporting.

Key Terms

average
bivariate analysis
codebook
contingency table
continuous variable
discrete variable
dispersion
frequency distribution
mean
median
mode
multivariate analysis
quantitative analysis
standard deviation
univariate analysis

Proposing Social Research: Quantitative Data Analysis

In this exercise, you should outline your plans for analysis. In earlier exercises, you'll have specified the variables to be analyzed, including precisely how you'll measure those variables.

Now you'll report how you plan to conduct your analysis. Are your aims primarily descriptive or explanatory? If explanatory, are you planning a simple bivariate analysis or a multivariate one? Here's where you should say whether you're planning a tabular analysis or something more complex than what has been discussed in this chapter. It doesn't really matter which computer program you use (SPSS, SAS, and so forth) unless it's a specialized program or one that is not commonly used.

If you've derived precise hypotheses, you may want to specify levels of statistical significance that will determine the meaning of the outcomes. This is not always necessary, however.

Review Questions

1. How might the various majors at your college be classified into categories? Create a coding system that would allow you to categorize them according to some meaningful variable. Then create a different coding system, using a different variable.
2. How many ways could you be described in numerical terms? What are some of your intrinsically numerical attributes? Could you express some of your qualitative attributes in quantitative terms?
3. How would you construct and interpret a contingency table from the following information: 150 Democrats favor raising the minimum wage, and 50 oppose it; 100 Republicans favor raising the minimum wage, and 300 oppose it?
4. Using the hypothetical data in the following table, how would you construct and interpret tables showing these three relationships?

    a. The bivariate relationship between age and attitude toward abortion

    b. The bivariate relationship between political orientation and attitude toward abortion

    c. The multivariate relationship linking age, political orientation, and attitude toward abortion

Age	Political Orientation	Attitude toward Abortion	Frequency
Young	Liberal	Favor	90
Young	Liberal	Oppose	10
Young	Conservative	Favor	60
Young	Conservative	Oppose	40
Old	Liberal	Favor	60
Old	Liberal	Oppose	40
Old	Conservative	Favor	20
Old	Conservative	Oppose	80

Online Study Resources

CENGAGENOW™

Go to
www.cengage.com/login

and click on "Create My Account" for access to this powerful online study tool. You'll get a personalized study plan based on your responses to a diagnostic pretest. Once you've mastered the material with the help of interactive learning tools, you can take a posttest to confirm that you're ready to move on to the next chapter.

Website for *The Basics of Social Research,* 5th edition

At the book companion website (www.cengage.com/sociology/babbie) you'll find many resources in addition to ***CengageNOW*** to aid you in studying for your exams. For example, you'll find Tutorial Quizzes with feedback, Internet Exercises, Flash Cards, Glossary and Crossword Puzzles, as well as Learning Objectives, GSS Data, Web Links, Essay Questions, and a Final Exam.