CHAPTER 1

CROSS-TABULATIONS

WHAT THIS CHAPTER IS ABOUT

In this chapter, we start with an introduction to the elements of quantitative analysis—the material to be covered in this book. Then we deal with the most basic of all quantitative analyst's tools, cross-tabulations or percentage tables. (Strictly speaking, not all percentage tables are cross-tabulations because we can percentage univariate distributions. But the main emphasis of this chapter will be on how to percentage tables involving the simultaneous tabulation of two or more variables.) Although the procedures are basic, they are not trivial. There are clear principles for deciding how to percentage cross-tabulations. We will cover these principles and also their exceptions. In the course of doing this, we will consider the logic of causal argument. Then we will consider other ways, besides percentage tables, of summarizing univariate and multivariate distributions of data, as well as ways of assessing the relative size of associations between pairs of variables *controlling for* or *holding constant* other variables. Take this chapter seriously, even if you have encountered percentage tables before and think you know a lot about them. In my experience, getting right the logic of how to percentage a table proves to be very difficult for many students, much more difficult than seemingly fancier procedures, such as multiple regression.

You will notice that many of the examples in the first three chapters are quite old, drawn from studies conducted as far back as the 1960s. This is because at that time tabular analysis was the "state of the art"—the technique used in most of the articles published in leading journals. Thus, by going back to the older research literature. I have been able to find particularly clear applications of tabular procedures.

INTRODUCTION TO THE BOOK VIA A CONCRETE EXAMPLE

In 1967, Gary Marx published an article in the *American Sociological Review* titled "Religion: opiate or inspiration of civil rights militancy among Negroes?" (Marx 1967a; see also Marx 1967b). The title expressed two competing ideas about how religiosity among Blacks might have affected their militancy regarding civil rights. One possibility was that religious people would be less militant than nonreligious people because religion gave them an other-worldly rather than this-worldly orientation, and established religious institutions have generally had a stake in the status quo and hence a conservative orientation. The other possibility was that they would be more militant because the Black churches were a major locus of civil rights militancy, and religion is an important source of universal humanistic values. Of course, a third possibility was that there would be no connection between religiosity and militancy.

Suppose that we want to decide which of these ideas is correct. How can we do this? One way—which is the focus of our interest here—would be to ask a probability sample of Blacks how religious they are and how militant with respect to civil rights they are, and then to cross-tabulate the answers to determine the relative likelihood, or probability, that religious and nonreligious people say they are militant. If religious people are *less* likely to give militant responses than are nonreligious people, the evidence would support the first possibility; if religious people are *more* likely to give militant responses, the evidence would favor the second possibility; and if there is no difference in the relative likelihoods of religious and nonreligious people giving militant responses, the evidence would favor the third possibility. Of course, evidence favoring an idea does not definitely prove it. I will say more about this later.

This seemingly simple example contains all of the elements that we will be dealing with in this book and that a researcher needs to take account of to arrive at a meaningful and believable answer to any research question. Let us consider the elements one by one.

First, the *idea*: is religion an opiate or inspiration of civil rights militancy? Without an idea, the manipulation of data is pointless. As you will see repeatedly, the nature of the idea a researcher wants to test will dictate the kind of data chosen and the manipulations performed. Without an idea, it is impossible to decide what to do, and the researcher will be tempted to try to do everything and be at a loss to choose from among the various things he or she has done. Ideas to be tested are generally called *hypotheses*; they also will be referred to here and in what follows as *theories*. A theory need not be either grandiose or abstract to be labeled as such. Any idea about what causes what, or why and how two variables are associated, is a theory.

Second is the information, or *data*, needed to test the idea or hypothesis (or theory). In this book, we will be concerned with data drawn from probability samples of populations. A *population* is any definable collection of things. Mostly we will be concerned with populations of people, such as the population of the United States. But social scientists are also interested in populations of organizations, cities, occupations, and so on. A *probability sample* is a subset of the population selected in such a way that the probability that a given individual in the population will be included in the sample is known. Only by using a probability sample is it possible to make inferences from the characteristics of the sample to the characteristics of the population from which the sample is drawn.

That is, if we observe a given result in a probability sample, we can infer within a specified range what the likely result will be in the population.

The sample used by Marx is actually quite complex, consisting of a probability sample of 492 Blacks living in metropolitan areas outside the South, plus four special samples: probability samples of Blacks living in Chicago, New York, Atlanta, and Birmingham. The total number of respondents from the non-Southern urban sample plus the four special samples is 1,119, and Marx treats the combined sample as representative of urban Blacks in the United States. This is not, in fact, entirely legitimate. Later we will explore ways to weight complex samples to make them truly representative of the populations from which they are drawn. Evaluation of the sample used in an analysis is an important part of the data analyst's task. But for now, we will go along with Marx in treating his sample as a probability sample of U.S. urban Blacks.

When our ideas are about the behavior or attitudes of people, a standard way of collecting data is to ask a probability sample chosen from an appropriate population to tell us about their behavior and attitudes by answering a set of specific questions. That is, we *survey* the sample by asking each individual in the sample a set of questions and recording the responses. In most sample surveys, the possible responses are preselected, and the person being surveyed, the *respondent*, is asked to choose the best response from a list (however, see the boxed comment on open-ended questions). For example, one of the questions Marx asked was

> *What would you say about the civil rights demonstrations over the last few years—that they have helped Negroes a great deal, helped a little, hurt a little, or hurt a great deal?*

Helped a great deal	1
Helped a little	2
Hurt a little	3
Hurt a great deal	4
Don't know	5

OPEN-ENDED QUESTIONS Occasionally, questions are worded in a way that requires a narrative response; these are known as *open-ended* questions. Open-ended questions are used when possible responses are too varied or complex to be conveniently listed on a questionnaire or when the researcher doesn't have a very good idea of what the possible responses will be. Open-ended questions must be *coded,* that is, converted into a standard set of response categories, as an editing operation in the course of data preparation. This is very time-consuming and expensive and is avoided whenever possible. Still, some items must be asked in an open-ended format. Both in the decennial census and in many contemporary surveys in the United States, for example, a series of three open-ended questions typically is asked to elicit information necessary to classify respondents according to standard detailed (three-digit) classifications of occupation and industry.

Each response, or *response category*, has a number associated with it, known as a code. The codes are what are actually recorded when the data are prepared for analysis because they are used to manipulate data in a computer. Typically, some respondents will refuse to answer a question or, in a self-administered questionnaire, will choose more than one response. Sometimes, an interviewer will forget to record a response or will record it in an ambiguous way. For these reasons, an extra code is usually designated to indicate nonresponses or uncodable responses. For example, code "9" might be assigned to nonresponses to the preceding question when the data are being prepared for analysis (this topic is discussed further a bit later). How to handle nonresponses, or missing data, is one of the perennial problems of the survey analyst, so we will devote a great deal of attention to this question.

The term *variable* refers to each set of response categories and the associated codes. A *machine-readable data set* (whether stored on computer tape, computer disk, floppy disks, CD-ROMs, thumb drives, or—almost extinct—IBM cards) consists of a set of codes for each individual in the sample corresponding to the response categories for the variables included in the data set. Suppose, for example, that the earlier question on whether civil rights demonstrations have helped Negroes is the tenth variable in a survey. Suppose, also, that the first respondent in the sample had said that demonstrations "helped a little." The data set would then include a "2" in the tenth location for the first individual. To know exactly what is included in a data set and where in the data set it is located, a *codebook* is prepared and used as a map to the data set. In Chapter Four, I will describe how to use a codebook. Here it is sufficient to note that the rudimentary materials necessary to carry out the sort of analysis dealt with in this book are a data set, a codebook for the data set, and documentation that describes the sample. We will not be concerned with problems of data collection or the preparation of a machine-readable data set, except in passing. These topics require full treatment in their own right, and we will not have time for them here.

It is customary to classify variables according to their level of measurement: nominal, ordinal, interval, or ratio. *Nominal variables* consist simply of a set of mutually exclusive and collectively exhaustive categories. Religious affiliation is an example of such a variable. For example, we might have the following response categories and codes:

Protestant	1
Catholic	2
Jewish	3
Other	4
None	5
No answer	9

Note that no order is implied among the responses—no response is "better" or "higher" than any other. The variable simply provides a way of classifying people into religious groups. Note, further, that every individual in the survey has a code, even those who didn't answer the question. This is accomplished by including a residual category, "Other," and a "No answer" category. In properly designed variables, categories are always mutually exclusive and collectively exhaustive—that is, written in such as way that each individual in the sample can be assigned one and only one code. (In Chapter Four, we will discuss various ways of coding missing data.)

Ordinal variables have an additional property—they can be arranged in an order along some dimension: quantity, value, or level. The question on civil rights demonstrations cited previously is an example of an ordinal variable, where the dimension on which the responses are ordered is helpfulness to Negroes. Actually, the variable is a useful example of what we often actually encounter in surveys. Two of the responses, "don't know" and the implicit "no answer" response, are not self-evidently ordered with respect to the other responses. In such situations, the analyst has two choices: either to exclude these responses from the analysis or to assign a position to them by recoding the variable, that is, altering the codes so that they indicate the new order. A plausible argument can be made that a "don't know" response is in between "helped a little" and "hurt a little," essentially a neutral rather than either a positive or a negative response. To treat the question in this way, an analyst would recode the variable by assigning code "3" to "don't know," code "4" to "hurt a little," and code "5" to "hurt a great deal." Whether to do this will depend on the specifics of the research question being investigated; but it is very important to be forthcoming about such manipulations when they are undertaken, reporting them as part of the writeup of the analysis. It would be rather more difficult to make the same sort of plausible case for including "no answer" as a neutral response because the bases for nonresponses are so varied, including simple error, failure to complete the questionnaire, and so on. Hence, there is no way to predict how nonrespondents would have responded had they done so. Therefore, it probably would be wisest to treat "no answer" as missing data.

The important feature of ordinal variables is that they include no information about the distance between categories. For example, we do not know whether the difference between a judgment that civil rights demonstrations "hurt a little" and the judgment that they "helped a little" is greater or smaller than the difference between the judgment that they "helped a little" and that they "helped a great deal." For this reason, some statisticians and social researchers argue that ordinal variables ought to be analyzed using ordinal statistics, which are statistics that make no assumptions about the distance between categories of a variable and use only the order property. This is not the position taken here. In this book, we will mainly consider two kinds of statistics, those appropriate for nominal variables and those appropriate for interval and ratio variables; the latter are known as parametric statistics. There are several reasons for ignoring statistics specifically designed for ordinal variables (with the exception of ordinal logistic regression, which we will consider in Chapter Fourteen). First, parametric statistics are much more powerful and far more mathematically tractable than ordinal statistics and, moreover, tend to be very robust; that is, they are generally quite insensitive to violations of assumptions about the nature of data—for example, that error is normally distributed. Second, ordinal statistics are much less widely used than parametric statistics; moreover, there are many alternatives for accomplishing the same thing and little consensus among researchers about which ordinal statistic to use. Third, many ordinal statistics involve implicit assumptions that are just as restrictive as the assumptions underlying parametric statistics. For example, it can be shown that Spearman's rank order correlation (an ordinal statistic) is identical to the product-moment (Pearson) correlation (the conventional parametric correlation coefficient) when interval or ratio variables are converted to ranks. In effect, then, the Spearman rank order correlation assumes an equal distance between each category rather than making no assumptions about the distance between categories. In sum, we gain little and lose much by using ordinal statistics. (However, if you are interested in such statistics, good discussions can be found in Davis 1971, and Hildebrand and others 1977.)

Interval variables and *ratio variables* are similar in that the distance between categories is meaningful. Not only can we say that one category is higher than another (on some dimension) but also how much higher. Such variables legitimately can be manipulated with standard arithmetic operations: addition, subtraction, multiplication, and division.

SAMUEL A. STOUFFER (1900–1960) was an early leader in the development of survey research. He was born in Sac City, Iowa, and earned a B.A. from Morningside College; earned an M.A. in literature at Harvard; served three years as an editor of the Sac City Sun, a newspaper founded by his father; and then began graduate studies in sociology at the University of Chicago, completing his Ph.D. in 1930. While at Chicago, he came under the tutelage of William F. Ogburn, who introduced him to statistics despite his self-described initial hostility to the subject. He studied statistical methods and mathematics intensively at Chicago and then spent a year as a Social Science Research Council Fellow at the University of London, where he worked with Karl Pearson, among others (see the biographical sketch of Pearson in Chapter Five). Stouffer held academic appointments in statistics and sociology at Wisconsin, Chicago, and Harvard. He was a skilled research administrator, heading a number of large projects designed to provide scientific understanding of major social crises: in the 1930s, a Social Science Research Council project to evaluate the influence of the Depression on social order, which resulted in thirteen monographs; during World War II, a study of soldiers for the Defense Department, which resulted in the classic publication, *The American Soldier* (Stouffer and others 1949); and in the 1950s, a study of the anticommunist hysteria of the McCarthy era, funded by the Ford Foundation's Fund for the Republic, which resulted in *Communism, Conformity, and Civil Liberties* (1955). When he died rather unexpectedly at age sixty after a brief illness, he was in the process of developing for the Population Council a new study on factors affecting fertility in developing nations. He also played an important role in developing the statistical program of the federal government, helping to establish the Division of Statistical Standards in the U.S. Bureau of the Budget. A hallmark of Stouffer's work is that he was strongly committed to using empirical data and quantitative methods to rigorously test ideas about social processes, which makes it fitting that a posthumous collection of his papers is titled *Social Research to Test Ideas* (1962).

Hence, we can compute statistics such as means and standard deviations for them. The difference between the two is that ratio variables have an intrinsic zero point, whereas interval variables do not. We can compare responses to ratio variables by taking the ratio of the value for one respondent (or group of respondents) to the value for another, whereas we can compare responses to interval variables only by taking the difference between them. Examples of interval variables include IQ and occupational prestige. Examples of ratio variables include years of school completed and annual income. It is not meaningful to say that someone's IQ is twice as high as someone else's, but it is meaningful to say that one person's IQ is 10 points higher than another person's IQ or that the within-race variance in IQ is larger than the between-race variance. By contrast, it is meaningful to say both that the incomes of men and women differ by $10,000 per year on the average and that the incomes of men are twice as high on average as those of women.

In this book, we often will treat ordinal variables as if they are interval variables to gain the power of parametric statistics. But we also will deal with procedures for assessing the adequacy of the interval assumption and for treating variables as nominal within the context of a general parametric approach that permits both nominal and interval or ratio variables to be dealt with simultaneously. These procedures involve various forms of regression analysis.

Often concepts of interest cannot be captured fully by single questions. For example, no single item in Marx's questionnaire fully captured what he meant by militancy. Hence, he constructed a multiple-item *scale* to represent this concept. Eight items that were pertinent to the situation in 1964 were used to construct a militancy scale. Individuals were classified as militant if they gave the militant response (shown in parentheses) to at least six of the eight items listed here (Marx 1967b, p. 41):

> *In your opinion, is the government in Washington pushing integration too slow, too fast, or about right?* (Too slow.)
>
> *Negroes who want to work hard can get ahead just as easily as anyone else.* (Disagree.)
>
> *Negroes should spend more time praying and less time demonstrating.* (Disagree.)
>
> *To tell the truth I would be afraid to take part in civil rights demonstrations.* (Disagree.)
>
> *Would you like to see more demonstrations or less demonstrations?* (More.)
>
> *A restaurant owner should not have to serve Negroes if he doesn't want to.* (Disagree.)
>
> *Before Negroes are given equal rights, they have to show that they deserve them.* (Disagree.)
>
> *An owner of property should not have to sell to Negroes if he doesn't want to.* (Disagree.)

There are many advantages to multiple-item scales, including in particular greater *reliability* and *validity* (both defined in Chapter Eleven). There also are many ways to construct multiple-item scales—some clearly superior to others—and some important pitfalls to avoid. Later—in Chapter Eleven—we will devote considerable attention to scale construction and evaluation.

The third element in any quantitative analysis is the *model*, the way we organize and manipulate data to assess our idea or hypothesis. The model has two components: the choice of statistical procedure and the assumptions we make about how the variables in our analysis are related. Given these two components, we can estimate the relative size or strength of the relationships between variables, and thus test our hypotheses (or ideas or theories) by assessing whether our estimates of the size of different effects are consistent with our hypotheses. For the simple example we have been considering, our models are cross-tabulations of militancy by religiosity (with the introduction of successive control variables, which are discussed a bit later), and our expectation (hypothesis) is that a higher percentage of the nonreligious than of the religious will be militant—or, because we have competing hypotheses, that a lower percentage of the nonreligious will be militant. Later in the book we will deal with statistical models that are more sophisticated—mostly variants of the general linear model—but the logic will remain unchanged. How we actually carry out cross-tabulation analysis is the topic of the next section.

CROSS-TABULATIONS

There are several ways to determine whether religious Blacks are more likely (or less likely) to be militant than are nonreligious Blacks. Perhaps the most straightforward approach is to cross-tabulate militancy by religiosity, that is, to count the frequency of persons with each combination of religiosity and militancy. By using four religiosity categories and two militancy categories, there are eight combinations of the two variables. In Marx's sample, the cross-tabulation of militancy by religiosity yields the following joint frequency distribution (Table 1.1).

TABLE 1.1. Joint Frequency Distribution of Militancy by Religiosity Among Urban Negroes in the U.S., 1964.

Religiosity	Militant	Nonmilitant	Total
Very religious	61	169	**230**
Somewhat religious	160	372	**532**
Not very religious	87	108	**195**
Not at all religious	25	11	**36**
Total	**333**	**660**	**993**

Source: Adapted from Marx (1967a, Table 6).

TECHNICAL POINTS ON TABLE 1.1

1) The Total row and Total column are known as marginals. They give the frequency distributions for each variable separately, in other words, the univariate frequency distributions. (Rows are read across and columns are read down.) The total number of cases (or respondents, or individuals) in the table is given in the lower-right cell (or position in the table). Note that this is fewer than the number of cases in the sample (recall that the sample consists of 1,119 cases). The difference is due to missing data; that is, some respondents did not answer all the questions needed to construct the religiosity and militancy scales. Later, we will deal extensively with missing data problems. For the present, however, we ignore the missing data and treat the sample as if it consists of 993 respondents.

2) The eight cells in the interior of the table give the bivariate frequency distribution, that is, the frequency of each combination of religiosity and militancy.

3) The titles of the variables and response categories are given in the table stubs.

4) When constructing a table, it is wise to check the accuracy of your entries by adding up the entries in each row and confirming that they correspond to the column marginal, for example, 61 + 169 = 230, and so on; adding up the entries of each column and confirming that they correspond to the row marginal, for example, 61 + 160 + 87 + 25 = 333, and so on; and adding up the row marginals and the column marginals and confirming that the sum of each corresponds to the table total. It is easy to introduce errors, especially when copying tables, and it is far better to discover them for yourself before committing them to print than for your readers to discover them after you have published. Always double-check your tables.

From this table, can we decide whether religiosity favors or inhibits militancy? Not very well. To do so, we would need to determine the *relative probability* that people of each degree of religiosity are militant. If the probability increases with religiosity, we would conclude that religiosity promotes militancy; if the probability of militancy decreases with religiosity, we would conclude that religion is an opiate. The relative probabilities are to be found by determining the conditional probability of militancy in each religiosity group, that is, the probability of militancy given that one is at a particular religiosity level. These conditional probabilities can be expressed as 61/230, 160/532, 87/195, and 25/36. Although this is a completely correct way of expressing the probabilities, they are more readily interpreted if expressed as percentages: (61/230)*100 = 27, and so on.

In fact, we ordinarily do this initially, by presenting tables of percentages rather than tables of frequencies. This makes direct comparisons of relative probabilities very easy. That is, we ordinarily would never present a table like Table 1.1 but instead would present a table like Table 1.2.

TABLE 1.2. **Percent Militant by Religiosity Among Urban Negroes in the U.S., 1964.**

Militancy	Very Religious	Somewhat Religious	Not Very Religious	Not at All Religious	Total
Militant	27%	30%	45%	69%	**33%**
Nonmilitant	73	70	55	31	**67**
Total	**100%**	**100%**	**100%**	**100%**	**100%**
N	(230)	(532)	(195)	(36)	(993)

Source: Table 1.1.

TECHNICAL POINTS ON TABLE 1.2

1) Always include the percentage totals (the row of 100%s). Although this may seem redundant and a waste of space, it makes it immediately clear to the reader in which direction you have percentaged the table. When the percentage totals are omitted, the reader may have to add up several rows or columns to figure it out. Using percentage signs on the top row of numbers and again on the Total row also clearly indicates to the reader that this is a percentage table.

2) Whole percentages are precise enough. There is no point in being more precise in the presentation of data than the accuracy of the data warrants. Moreover, fractions of percentages are usually uninteresting. It is hard to imagine anyone wanting to know that 37.44 percent of women and 41.87 percent of men do something; it is sufficient to note that 37 percent of women and 42 percent of men do it. Incidentally, a convenient rounding rule is to round to the even number. Thus, 37.50 becomes 38, but 36.50 becomes 36. Of course, 36.51 becomes 37 and 37.49 also becomes 37. You only want to report more than whole percentages if you have a distribution with many categories and are concerned about rounding error.

3) Always include the number of cases on which the percentages are based (that is, the denominator for the percentages). This enables the reader to reconstruct the entire table of frequencies (within the limits of rounding error) and hence to reorganize the data into a different form. Note that Table 1.2 contains all of the information that Table 1.1 contains because you can reconstruct Table 1.1 from Table 1.2: 27 percent of 230 is 62.1, which rounds to 62 (within rounding error of 61), and so on. Customarily, percentage bases are placed in parentheses to clearly identify them and to help them stand out from the remainder of the table.

4) Sometimes it is useful to include a Total column, as I have done here, and sometimes not. The choice should be based on substantive considerations. In the present case, about one-third of the total sample is militant (as defined by Marx); hence, the marginal distribution for the dependent variable is reported here. Recall from page 7 that "militants" are those who gave militant responses to at least six of the eight items in the militancy scale. We now see that about one-third of the sample did so. Obviously, if we defined as militant all those who gave at least five militant responses, the percentage militant would be higher.

5) No convention dictates that tables must be arranged so that the percentages run down, that is, so that each column totals to 100 percent. In Table 1.2, the categories of the dependent variable form the rows, and the categories of the independent variable form the columns. If it is more convenient to reverse this, so that the categories of the independent variable form the rows, this is perfectly acceptable. The only caveat is that within each category of the independent variable, the percentage distribution across the categories of the dependent variable must total to 100 percent. Thus, if the categories of the dependent variable form the columns, the table should be percentaged across each row.

The Direction to Percentage the Table

Note that the direction in which this table is percentaged is not at all arbitrary but rather is determined by the nature of the hypothesis being tested. The question being addressed is whether religiosity promotes or hinders militancy. In this formulation, religiosity is presumed to influence, cause, or determine militancy, not the other way around. (One could imagine a hypothesis that assumed the opposite—we might suppose that militants would tend to lose interest in religion as their civil rights involvement consumed their passions. But that is not the idea being tested here.) The variable being determined, influenced, or caused is known as the *dependent* variable, and the variables that are doing the causing, determining, or influencing are known as *independent*, or *predictor*, variables. The choice of causal order is always a matter of theory and cannot be determined from the data.

The choice of causal order then dictates the way the table is constructed. Tables should (almost—an exception will be presented later) always be constructed to express the conditional probability of being in each of the categories of the dependent variable given that an individual is in a particular category of the independent variable(s). (Do not let the fact that the table is expressed in percentages and the rule is expressed in probabilities confuse you. A percentage, which means "per hundred," is just a probability multiplied by 100. Percentages range from 0 to 100; probabilities range from 0 to 1.00.) Thus, in Table 1.2, I show the percentage militant for each religiosity category; that is I show the conditional probability (×100) of being militant, given that an urban Black is, respectively, very religious, somewhat religious, not very religious, or not at all religious. Note that the probability of being militant increases as religiosity decreases. Of the very religious, 27 percent are militant, as are 30 percent of the somewhat religious, 45 percent of the not very religious, and 69 percent of the not at all religious. Thus, given the formulation with which I (and Marx) started, in which religiosity was posited as alternatively an opiate or an inspiration, we are led to conclude that religiosity is an opiate because the more religious people are, the less likely they are to be militant.

It is important to understand this example thoroughly because the logic of which way to compute percentages and which comparisons to make is the same in all cross-tabulation tables.

Control Variables

Thus far, we have determined that the probability of militancy increases as religiosity decreases. Do we want to stop here? To do so would be to accept religiosity as the causative factor, that is, to conclude that religiosity causes people to be less militant. If we had a strong theory that predicted an inverse relationship between religiosity and militancy, regardless of anything else, we might be prepared to accept our two-variable cross-tabulation as an adequate test. Ordinarily, however, we will want to consider whether there are alternative explanations for the relationships we observe. In the present instance, for example, we might suspect that both religiosity and militancy are determined by some third factor. One obvious possibility is education. We might expect well-educated Blacks to be both less religious and more militant than more poorly educated Blacks. If this is so, religiosity and militancy would appear to be inversely related even if there were no causal connection between them. This is known as a *spurious association* or spurious correlation.

How can we test this possibility?

First, we need to determine whether education does in fact reduce religiosity by creating Table 1.3. This table shows that among urban Blacks in 1964, those who are well educated tend to be less religious. Of those with only a grammar school education, 31 percent are very religious, compared to 19 percent of those with a high school or college education. Further, only 1 percent of those with a grammar school education, 4 percent of those with a high school education, and 11 percent of those with a college education are not at all religious. Thus, we can say that education and religiosity are inversely or negatively associated: as education increases religiosity decreases. (Study this table carefully to see why it is percentaged as it is. What would you be asserting if you percentaged the table in the other direction?)

Next we need to determine whether education increases militancy by creating Table 1.4.

From Table 1.4, we see that the higher the level of educational attainment, the greater the percentage militant. Only 22 percent of those with grammar school education, 36 percent of those with high school education, and fully 53 percent of those with college education are militant. Another way of putting this is to say that a positive association exists between education and militancy: as education increases, the probability of militancy increases.

TABLE 1.3. Percentage Distribution of Religiosity by Educational Attainment, Urban Negroes in the U.S., 1964.

	Educational Attainment		
Religiosity	**Grammar School**	**High School**	**College**
Very religious	31%	19%	19%
Somewhat religious	57	54	45
Not very religious	12	24	25
Not at all religious	1	4	11
Total	**101%**	**101%**	**100%**
N	(353)	(504)	(136)

Source: Adapted from Marx (1967a, Table 6).

TABLE 1.4. Percent Militant by Educational Attainment, Urban Negroes in the U.S., 1964.

	Educational Attainment		
Militancy	**Grammar School**	**High School**	**College**
Militant	22%	36%	53%
Nonmilitant	78	64	47
Total	**100%**	**100%**	**100%**
N	(353)	(504)	(136)

Source: Adapted from Marx (1967a, Table 6).

TECHNICAL POINTS ON TABLE 1.3

1) Sometimes your percentages will not total to exactly 100 percent due to rounding error. Deviations of one percentage point (99 to 101) are acceptable. Larger deviations probably indicate computational error and should be carefully checked.

2) Note how the title is constructed. It states what the table is (a percentage distribution), which variables are included (the convention is to list the dependent variable first), what the sample is (urban Negroes in the U.S.), and the date of data collection (1964). The table should always contain sufficient information to enable one to read it without referring to the text. Thus, the title and variable headings should be clear and complete; if there is insufficient space to do this, it should be done in footnotes to the table.

3) In the interpretation of percentage distributions, comparing the extreme categories and ignoring the middle categories is usually sufficient. Thus, we noted that the proportion "very religious" decreases with education, and the proportion "not at all religious" increases with education. Similar assertions about how the middle categories ("somewhat religious" and "not very religious") vary with education are awkward because they may draw from or contribute to categories on either side. For example, the percentage "not very religious" among those with a college education might be larger if either the percentage "somewhat religious" or the percentage "not at all religious" were smaller. But one shift would indicate a more religious college-educated population, and the other shift would indicate a less religious college-educated population. Hence, the "not very religious" row cannot be interpreted alone, and usually little is said about the interior rows of a table. On the other hand, it is important to present the data so that the reader can see that you have not masked important details and to allow the reader to reorganize the table by collapsing categories (discussed later).

4) In dealing with scaled variables, such as religiosity, you should not make much of the relative size of the percentages within each distribution; that is, comparisons should be made across the categories of the independent variable, not across the categories of the dependent variable. In the present case, it is legitimate to note that those with a grammar school education are more likely to be very religious than are those who are better educated, but it is not legitimate to assert that more than half those with a grammar school education are somewhat religious. The reason for this is that the scale is only an ordinal scale; the categories do not carry an absolute value. How religious is "very religious"? All we know is that it is more religious than "somewhat religious." In consequence, it is easy to change the distribution simply by combining categories. Suppose, for example, we summed the top two rows and called the resulting category "religious." In this case, 88 percent of those with grammar school education would be shown as "religious." Consider how this would change the assertions we would make about this sample if we took the category labels seriously.

TECHNICAL POINTS ON TABLE 1.4

1) When you are presenting several tables involving the same data, always check the consistency of your tables by comparing numbers across the tables wherever possible. For example, the number of cases in Table 1.4 should be identical to that in Table 1.3.

Because educated urban Blacks are both less likely to be religious and more likely to be militant than are their less educated counterparts, it is possible that the observed association between religiosity and (non)militancy is determined entirely by their mutual dependence on education and that there is no connection between militancy and religiosity among people who are equally well educated. If this proves true, we would say that education *explains* the association between religiosity and militancy and that the association is spurious because it does not arise from a causal connection between the variables.

To test this possibility, we study the relation between militancy and religiosity within categories of education by creating a three-variable cross-tabulation of militancy by religiosity by education. Such a table can be set up in two different ways. The first is shown in Table 1.5, and the second in Table 1.6.

TABLE 1.5. Percent Militant by Religiosity and Educational Attainment, Urban Negroes in the U.S., 1964.

	Grammar School			High School			College		
Militancy	**V**	**S**	**N**	**V**	**S**	**N**	**V**	**S**	**N**
Militant	17%	22%	32%	34%	32%	47%	38%	48%	68%
Nonmilitant	83	78	68	66	68	53	62	52	32
Total	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
N	(108)	(201)	(44)	(96)	(270)	(138)	(26)	(61)	(49)

Source: Adapted from Marx (1967a, Table 6).
*V=very religious; S=somewhat religious; N=not very religious or not at all religious.

TECHNICAL POINTS ON TABLE 1.5

1) In this sort of table, education is the control variable. The table is set up to show the relationship between militancy and religiosity within categories of education, that is (synonymously), "controlling for education," "holding education constant," or "net of education." The control variable should always be put on the outside of the tabulation so that it changes most slowly. This format facilitates reading the table because it puts the numbers being compared in adjacent columns. (Sometimes we want to study the relationship of each of two independent variables to a dependent variable, in each case controlling for the other. In such cases, we still make only one table and construct it in whatever way made it easiest to read. If our dependent variable is dichotomous or can be treated as dichotomous, we set up the table in the format of Table 1.6.)

2) Note that the "not very religious" and "not at all religious" categories were combined. This is often referred to as collapsing categories. Collapsing is usually done when there would be too few cases to produce reliable results for some categories. In the present case, as we know from Table 1.1 or 1.2, there are thirty-six people who are not at all religious. Dividing them on the basis of educational attainment would produce too few cases in each group to permit reliable estimates of the percent militant. Hence, they were combined with the adjacent group, "not very religious."

   An additional reason for collapsing categories is to improve clarity. Too much detail makes it difficult for the reader to grasp the main features of the table. Often, it helps to reduce the number of categories presented. On the other hand, if categories of the independent variable differ in terms of their distribution on the dependent variable, combining the categories will mask important distinctions. A fine balance must be struck between clarity and precision, which is why constructing tables is an art.

From Table 1.5, we see that religiosity continues to inhibit militancy even when education is controlled, although the differences in percent militant among religiosity categories tend to be smaller than in Table 1.2 where education is not controlled. (In the next chapter, we will discuss a procedure for calculating the size of the reduction in an association resulting from the introduction of a control variable, the *weighted net percentage difference*.) Among those with grammar school education, 17 percent of the very religious and 32 percent of the not religious are militant; the corresponding percentages for those with high school education are 34 and 47 and for those with college education are 38 and 68. Thus we conclude that education does not completely account for the inverse association between religiosity and militancy.

At this point, we have to decide whether to continue the search for additional explanatory variables. Our decision usually will be based on a combination of substantive and technical considerations. If we have grounds for believing that some other factor might account both for religiosity and militancy, net of education, we probably would want to control for that factor as well. Note, however, that the power of additional factors to explain the association between two original variables (here religiosity and militancy) will depend on their association with previously introduced control variables. To the extent that additional variables are highly correlated with variables already introduced, they will have little impact on the association. This is an extremely important point that will recur in the context of multiple regression analysis. Be sure you understand it thoroughly.

Consider age. What relation would you expect age to have to religiosity and to militancy?

Pause to Think About This

Religiosity is likely positively associated with age—that is, older people tend to be more religious—and militancy is inversely associated with age—younger people tend to be more militant. Hence, we might expect the association between religiosity and militancy to be a spurious function of age. That is, within age categories, there may be no association between religiosity and militancy.

What, however, of the relation between age and education? In fact, from knowledge about the secular trend in education among Blacks, we would expect younger Blacks to be substantially better educated than older Blacks. To the extent this is true, age and education are likely to have similar effects on the association between religiosity and militancy. Hence, introducing age as a control variable in addition to education is not likely to reduce the association between religiosity and militancy by much, relative to the effect of education alone.

Apart from theoretical and logical considerations (is a variable theoretically relevant, and is it going to add anything to the explanation?), there is a straightforward technical reason for limiting the number of variables included in a single cross-tabulation—we quickly run out of cases. Most sample surveys include a few hundred to a few thousand cases. We already have seen that a three-variable cross-tabulation required that we collapse two of the religiosity categories. A four-variable cross-tabulation of the same data is likely to yield so many small percentage bases as to make the results extremely unreliable. The difficulty in studying more than about three variables at a time in a cross-tabulation provides a strong motivation to use some form of regression analysis instead. A substantial fraction of the chapters to follow will be devoted to the elaboration of regression-based procedures.

Table 1.5 also enables us to assess the effect of education on militancy, controlling for religiosity by comparing corresponding columns in each of the three panels. Thus, we note that, among those who are very religious, 17 percent of the grammar school educated are militant compared to 34 percent of the high school educated and 38 percent of the college educated; among those who are somewhat religious, the corresponding percentages are 22, 32, and 48; and among those who are not religious, they are 32, 47, and 68. Hence, we conclude that, at any given level of religiosity, the better educated are more militant.

TABLE 1.6. Percent Militant by Religiosity and Educational Attainment, Urban Negroes in the U.S., 1964 (Three-Dimensional Format).

	Educational Attainment		
Religiosity	Grammar School	High School	College
Very religious	17% (108)	34% (96)	38% (26)
Somewhat religious	22% (201)	32% (270)	48% (61)
Not very or not at all religious	32% (44)	47% (138)	68% (49)

Source: Table 1.5.

TECHNICAL POINTS ON TABLE 1.6

1) Each pair of entries gives the percentage of people who have a trait and the percentage base, or denominator, of the ratio from which the percentage was computed. Thus, the entry in the upper-left corner indicates that 17 percent of the 108 very religious grammar-school-educated people in the sample are militant. From this table, we can reconstruct any of the preceding five tables (but with the two least religious categories collapsed into one), within the limits of rounding error. Try to do this to confirm that you understand the relationships among these tables.

This requires a fairly tedious comparison, however, skipping around the table to locate the appropriate cells. When the dependent variable is dichotomous, that is, has only two response categories, a much more succinct table format is possible and is preferred. Table 1.6 contains exactly the same information as Table 1.5, but the information is arranged in a more succinct way. Tables like Table 1.6 are known as three-dimensional tables.

Compare Tables 1.5 and 1.6. You will see that they contain exactly the same information—all the additional numbers in Table 1.5 are redundant. Moreover, Table 1.6 is much easier to read because we can see the effect of religiosity on militancy, holding constant education, simply by reading down the columns, and can see the effect of education on militancy, holding constant religiosity, simply by reading across the rows.

WHAT THIS CHAPTER HAS SHOWN

In this chapter, we have seen an initial idea formulated into a research problem, an appropriate sample chosen, a survey conducted, and a set of variables created and combined into scales to represent the concepts of interest to the researcher. We then considered how to construct a percentage table that shows the relationship between two variables, with special attention to determining in which direction to percentage tables using the concept of conditional probability distributions—the probability distribution over categories of the dependent variable computed separately for each category of the independent variable(s). This is the most difficult concept in the chapter, and one you should make sure you completely understand.

The other important concept you need to understand fully is the idea of statistical controls, also known as controlling for or holding constant confounding variables, to determine whether relationships hold within categories of the control variable(s). Finally, we considered various technical issues regarding the construction and presentation of tables. The aim of the game is to construct attractive, easy to read tables.

In the next chapter, we continue our discussion of cross-tabulations, considering various ways of analyzing tables with more than two variables and, more generally, the logic of multivariate analysis.

CHAPTER 2

MORE ON TABLES

WHAT THIS CHAPTER IS ABOUT

In this chapter we expand our understanding of how to deal with cross-tabulations, both substantively and technically. First we continue our consideration of the *logic of elaboration*, that is, the introduction of additional variables to an analysis; second, we consider a special situation known as a *suppressor* effect, when the influences of two independent variables offset each other; third, we consider how variables combine to produce particular effects, drawing a distinction between *additive* and *interaction* effects; fourth, we see how to assess the effect of a single independent variable in a multivariate percentage table while controlling for the effects of the other independent variables via *direct standardization*; and finally we consider the distinction between *experiments* and *statistical controls*.

THE LOGIC OF ELABORATION

In traditional treatments of survey research methods (for example, Lazarsfeld 1955; Zeisel 1985), it was customary to make a distinction between two situations in which a third variable completely or partially accounts for the association between two other variables: *spurious* associations and associations that can be accounted for by an *intervening* variable or variables. The distinction between the two is that when a control variable (Z) is temporally or causally prior to an independent variable (X) and dependent variable (Y), and when the control variable completely or partly explains the association between the independent and dependent variable, we infer that there is no causal connection or only a weak causal connection between the independent and dependent variables. However, when the control variable intervenes temporally or causally between the independent and dependent variables, we would not claim that there is no causal relationship between the independent and dependent variables but rather that the intervening variable explains, or helps explain, how the independent variable exerts its effect on the dependent variable. In the previous chapter we considered spurious associations. In this chapter we revisit spurious associations and also consider the effect of intervening variables.

PAUL LAZARSFELD (1901–1976) was a major force—perhaps the preeminent figure—in the establishment of survey research as the dominant method of data collection in American sociology. Born in Vienna, he earned a doctorate in mathematics (with a dissertation that dealt with mathematical aspects of Einstein's gravitational theory). He came to sociology by way of his 1926 marriage to Marie Jahoda and joined with her (and Hans Zeisel) in the now-classic Marienthal study (see the biosketch of Zeisel later in this chapter). His interest in social research lasted, although his marriage to Jahoda did not. He came to the United States in 1933, with an appointment first at the University of Newark (now part of Rutgers University) and then at Columbia University, where he founded the Bureau for Applied Social Research, the training ground for many prominent social scientists and the organizational base for many innovations in quantitative data collection and analysis. One of the hallmarks of Lazarsfeld's approach was that social research was most effectively accomplished as a collective endeavor, involving a team of specialists. But perhaps most important was his insistence that tackling applied questions was a legitimate endeavor of academic sociology and that answers to such questions could contribute to the theoretical development of the discipline.

Spurious Association

Consider the three variables, X, Y, and Z. Suppose that you had observed an association between X and Y and suspected that it might be completely explained by the dependence of both X and Y on Z. (For a substantive example, recall the hypothesis in the previous chapter that the negative relation between religiosity and militancy was due to the dependence of both on education—Blacks with more education were both less religious and more militant.) Such a hypothesis might be diagrammed as shown in Figure 2.1.

Causal diagrams of this sort are used for purposes of explication throughout the book. They are extensively used in *path analysis*, which is a way of representing and algebraically manipulating structural equation models that was widely used in the 1970s but is less frequently encountered now (see additional discussion of structural equation models and path analysis in Chapter Sixteen). My use of such models is purely heuristic. Nonetheless, I use them in such a way as to be conceptually complete. Hence, the paths from x to X (p_{Xx}) and from y to Y (p_{Yy}) indicate that other factors besides Z influence X and Y.

Now, if the association between X and Y within categories of Z were very small or nonexistent, we would regard the association between X and Y as entirely explained by their mutual dependence on Z. However, this generally does not happen; recall, for example, that the negative association between religiosity and militancy did not disappear when education was held constant. We ordinarily do¡ not restrict ourselves to an all-or-nothing hypothesis of spuriousness—except in the exceptional case where we have a very strong theory requiring that a particular relation be completely spurious; rather, we ask what the association is between X and Y controlling for Z (and what the association is between Z and Y controlling for X). The logic of our analysis can be diagramed as shown in Figure 2.2.

To state the same point differently, rather than assuming that the causal connection between X and Y is zero and determining whether our assumption is correct, we estimate the relation between X and Y holding constant Z and determine its size—which, of course, may be zero, in which case Figure 2.1 and Figure 2.2 are identical.

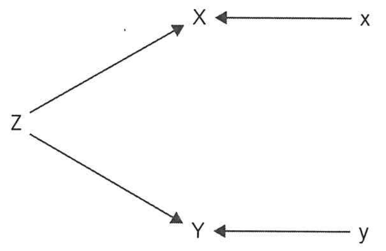

FIGURE 2.1. *The Observed Association Between X and Y Is Entirely Spurious and Goes to Zero When Z Is Controlled.*

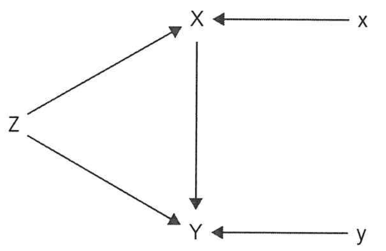

FIGURE 2.2. *The Observed Association Between X and Y Is Partly Spurious: the Effect of X on Y Is Reduced When Z Is Controlled (Z Affects X and Both Z and X Affect Y).*

Intervening Variables

Now let us consider the intervening variable case. Suppose we think two variables, *X* and *Y*, are associated only because *X* causes *Z* and *Z* causes *Y*. An example might be the relation between a father's occupation, son's education, and son's income. Suppose we expect the two-variable association between *X* and *Y*—sometimes called the *zero-order association*, short for zero-order partial association, that is, no partial association—to be positive, but think that this is due entirely to the fact that the father's occupational status influences the son's education and that the son's education influences the son's income; we think there is no direct influence of the father's occupational status on the son's income, only the indirect influence through the son's education. This sort of claim can be diagrammed as shown in Figure 2.3.

But, as before, unless we have a very strong theory that depends on there being no direct connection between *X* and *Y*, we probably would inspect the data to determine the influence of *X* on *Y*, holding constant the intervening variable *Z*, and would also determine the influence of *Z* on *Y*, holding constant the antecedent variable *X*. This can be diagrammed as shown in Figure 2.4

If the net, or *partial*, association between *X* and *Y* proves to be zero, we would conclude that a chain model of the kind described in Figure 2.3 describes the data. Otherwise, we would simply assess the strength and nature of both associations, between *X* and *Y* and between *Z* and *Y* (and, for completeness, the zero-order association between *X* and *Z*).

Notice the similarity between Figure 2.2 and Figure 2.4. With respect to the ultimate dependent variable, *Y*, the two models are identical. The only difference has to do with the specification that *Z* causes *X* or that *X* causes *Z*. There is still another possibility: *X* and *Z* cause *Y*, but no claim is made regarding the causal relation between *X* and *Z*. This can be diagrammed as shown in Figure 2.5.

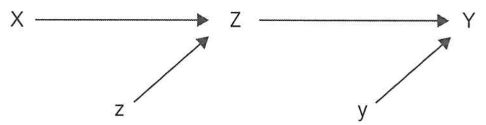

FIGURE 2.3. *The Observed Association Between X and Y Is Entirely Explained by the Intervening Variable Z and Goes to Zero When Z Is Controlled.*

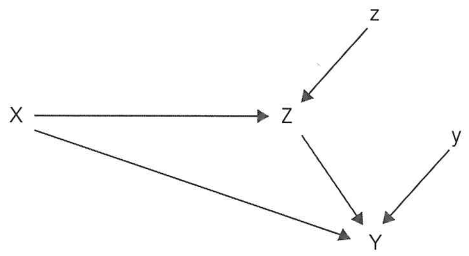

FIGURE 2.4. *The Observed Association Between X and Y Is Partly Explained by the Intervening Variable Z: the Effect of X on Y Is Reduced When Z is Controlled (X Affects Z, and Both X and Z Affect Y).*

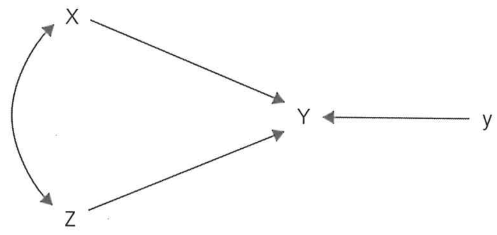

FIGURE 2.5. *Both X and Z Affect Y, but there Is no Assumption Regarding the Causal Ordering of X and Z.*

In almost all of the analyses we undertake—including cross-tabulations of the kind we are concerned with at present, multivariate models in an ordinary least squares regression framework, and log-linear and logistic analogs to regression for categorical dependent variables—the models, or theories, represented by Figures 2.2, 2.4, and 2.5 will be analytically indistinguishable with respect to the dependent variable, *Y*. The distinction among them thus must rest with the ideas of the researcher, not with the data. From the standpoint of data manipulation, all three models require assessing the net effect of each of two variables on a third variable, that is, the effect of each independent variable holding constant the other independent variable. Obviously, the same ideas can be generalized to situations involving more than three variables.

SUPPRESSOR VARIABLES

One final idea needs to be discussed here, the notion of *suppressor variables*. Thus far, we have dealt with situations in which we suspected that an observed association between two variables was due to the effect of a third, either as an antecedent or an intervening variable. Situations can arise, however, in which there appears to be *no* association between two variables when, in fact, there is a causal connection. This happens when some other variable is related to the two variables in such a way that it *suppresses* the observed zero-order association—specifically, when one independent variable has opposite effects on another independent variable and on the dependent variable, and the two independent variables have opposite effects on the dependent variable. Such situations can be diagrammed as shown in Figure 2.6.

For example, suppose you are interested in the relations among education, income, and fertility. On theoretical grounds, you might expect the following: education will have a positive effect on income; holding constant income, education will have a negative effect on fertility (the idea being that educated people want to do more for their children and regard children as more expensive than do poorly educated people; hence at any given level of income, they have fewer children); holding constant education, the higher the income, the higher the number of children (the idea being that children are generally regarded as desirable so that at any given level of the perceived cost of children, those with more to spend, that is, with higher income, will have more children). These relationship are represented in Figure 2.6, where *X* = level of education, *Z* = income, and *Y* = number of

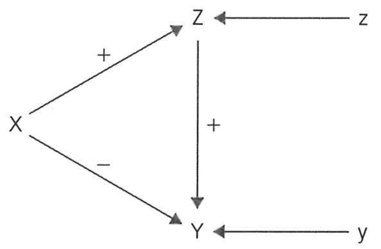

FIGURE 2.6. *The Size of the Zero-Order Association Between X and Y (and Between Z and Y) Is Suppressed When the Effects of X on Z and Y have Opposite Sign, and the Effects of X and Z on Y have Opposite Sign.*

children. The interesting thing about this diagram is that it implies that the gross, or zero-order, relationships between *X* and *Y* and between *Z* and *Y* will be smaller than the net, or first-order partial, relationships and might even be zero, depending on the relative size of the associations among the three variables. To see how this happens, consider the relationship between education and fertility. We have posited that educated people tend to have more income, and at any given level of education, higher-income people tend to have more children. Hence, so far, the relation between education and fertility would be expected to be positive. But we also have posited that at any given income level, better-educated people tend to have fewer children. So we have a positive causal path and a negative causal path at work at once, and the effect of each is to offset or *suppress* the other effect, so that the zero-order relationship between education and fertility is reduced.

ADDITIVE AND INTERACTION EFFECTS

We now consider *interaction effects*, situations in which the effect of one variable on another is contingent on the value of a third variable. To see this clearly, consider Table 2.1.

This table shows that in 1965 educational attainment had no effect on acceptance of abortion among Catholics but that among Protestants, the greater the education, the greater the percentage accepting abortion. Thus, Catholics and Protestants with 8th grade education or less were about equally likely to believe that legal abortion should be permitted under specified circumstances but, among those with more education, Protestants were substantially more likely to accept abortion than were Catholics. Among the college educated, the difference between the religious groups is fully twenty points: about 31 percent of Catholics and 51 percent of Protestants believed that abortion should be permitted.

This kind of result is called an *interaction effect*. Religion and educational attainment *interact* to produce a result different from what each would produce alone. That is, the relationship between education and acceptance of abortion differs for Catholics and Protestants, and the relationship between religion and acceptance of abortion differs by education. Situations in which the relationship between two variables depends on the value of a third, as it does here, are known as interactions. In the older survey analysis literature (for example, Lazarsfeld 1955; Zeisel 1985), interactions are sometimes called *specifications*. Religion *specifies* the relationship between education and beliefs about abortion: acceptance of abortion increases with education among Protestants but not among Catholics.

TABLE 2.1. **Percentage Who Believe Legal Abortions Should Be Possible Under Specified Circumstances, by Religion and Education, U.S. 1965 (N = 1,368; Cell Frequencies in Parentheses).**

	Educational Attainment			
Religion	8th Grade or Less	Some High School	High School Graduate	Some College or More
Catholic	31% (90)	33% (96)	33% (89)	31% (75)
Protestant	29% (287)	36% (250)	43% (256)	51% (225)

Source: Rossi (1966).
Note: Non-Christians omitted.

HANS ZEISEL (1905–1992) is best known for his classic book on how to present statistical tables, *Say It with Figures*, which has appeared in six editions and has been translated into seven languages. Born in what is now the Czech Republic, Zeisel was raised and educated in Vienna, where he and Paul Lazarsfeld were personal and professional collaborators—both leaders of the Young Socialists and, together with Marie Jahoda, authors of the now classic study of the social impact of unemployment on Marienthal, a small Austrian community (1971 [1933]). Zeisel earned a law degree from the University of Vienna in 1927 and practiced law there as well as conducting social research until 1938, when the *Anschluss* forced him to flee to New York. There he carried out market research at McCann-Erickson and then the Tea Council until 1953, when he joined the law faculty at the University of Chicago, where he conducted a number of empirical studies in the area now known as "law and society" research. He was a pioneer in the use of statistical evidence and survey data in legal cases, often serving as an expert witness.

Where we do not have interaction effects we have *additive effects* (or no effects). Suppose, for example, that instead of the numbers in Table 2.1, we had the numbers shown in Table 2.2.

What would this show? We could say two things: (1) The effect of religion on acceptance of abortion is the same at all levels of education. That is, the difference between the percentage of Catholics and Protestants who would permit abortion is 10 percent at each level of education. (2) The effect of education on acceptance of abortion is the same for Catholics and Protestants. For example, the difference between the percentage who would permit abortion among those with some high school education and those who are high school graduates is 10 percent for both Catholics and Protestants. Similarly, the difference between the percentage who would permit abortion among those with a high school education and among those with at least some college is 20 percent for both Catholics and Protestants.

TABLE 2.2. Percentage Accepting Abortion by Religion and Education (Hypothetical Data).

	8th Grade or Less	Some High School	High School Graduate	Some College or More
Catholic	30%	35%	45%	65%
Protestant	40%	45%	55%	75%

The reason this table is *additive* is that the effects of each variable add together to produce the final result. It is as if the probability of any individual in the sample accepting abortion is at least .3 (so we could add .3 to every cell in the table); the probability of a Protestant accepting abortion is .1 greater than the probability of a Catholic accepting abortion (so we could add .1 to all the cells containing Protestants); the probability of someone with some high school accepting abortion is .05 greater than the probability of someone with an 8th grade education doing so (so we would add .05 to the cells for those with some high school); the probability of those who are high school graduates accepting abortion is .15 greater than the probability of those with an 8th grade education doing so (so we would add .15 to the cells for those with high school degrees); and the probability of those with some college accepting abortion is .35 higher than the probability of those with an 8th grade education doing so (so we would add .35 to the cells for those with at least some college). This would produce the results we see in Table 2.2 (after we convert proportions to percentages by multiplying each number by 100).

By contrast, it is not possible to add up the effect of each variable in a table containing interactions because the effect of each variable depends on the value of the other independent variable or variables.

Many relationships of interest to social scientists involve interactions—especially with gender and to some extent with race; but it is also true that many relationships are additive. Adequate theoretical work has not yet been done to allow us to specify very well in advance which relationships we would expect to be additive and which relationships we would expect to involve interactions.

Later you will see more sophisticated ways to distinguish additive effects from interactions and to deal with various kinds of interactions via log-linear analysis and regression analysis.

DIRECT STANDARDIZATION

Often we want to assess the relationship between two variables controlling for additional variables. Although we have seen how to assess partial relationships—that is, relationships between two variables within categories of one or several control variables—it would be helpful to have a way of constructing a single table that shows the average relationship between two variables *net of*, that is, *controlling for*, the effects of other variables. *Direct standardization* provides a way of doing this. Note that this technique has other names, for example, *covariate adjustment*. However, the technique is most widely used in demographic research, so I use the term by which it is known in demography, *direct standardization*. It is important to understand that, even though the same term is used, this procedure has no relationship to standardizing variables to create a common metric. We will consider this subject in Chapter Five.

Example 1: Religiosity by Militancy Among U.S. Urban Blacks

The procedure is most easily explained in the context of a concrete example. Thus we revisit the analysis shown in Tables 1.2 through 1.6 of Chapter One (slightly modified). Recall that we were interested in whether the relationship between militancy and religiosity among Blacks in the United States could be explained by the fact that better-educated Blacks tend to be both less religious and more militant. Because education does not completely explain the association between militancy and religiosity, it would be useful to have a way of showing the association remaining after the effect of education has been removed. We can do this by getting an adjusted percentage militant for each religiosity category, which we do by computing a weighted average of the percent militant across education categories within each religion category but with the weights taken from the overall frequency distribution of education in the sample. (Alternatively, because they are mathematically identical, we can compute the weighted *sum*, using as weights the *proportion* of cases in each category.) By doing this, we construct a hypothetical table showing what the relationship between religiosity and militancy would be if all religiosity groups had the same distribution of education. It is in this precise sense that we can say we are showing the association between religiosity and militancy net of the effect of education. As noted earlier, this procedure is known as direct standardization or covariate adjustment.

Note that the weights need not be constructed from the overall distribution in the table. Any other set of weights could be applied as well. For example, if we wanted to assess the association between religiosity and militancy on the assumption that Blacks had the same distribution of education as Whites, we would treat Whites as the *standard population* and use the White distribution across educational categories (derived from some external source) as the weights. We will see two examples of this strategy a bit later in the chapter.

Now let us construct a militancy-by-religiosity table adjusted, or standardized, for education, to see how the procedure works. We do this from the data in Table 1.6. First, we derive the standard distribution, the overall distribution of education. Because there are 993 cases in the table (= 108 + . . . + 49), and there are 353 (= 108 + 201 + 44) people with a grammar school education, the proportion with a grammar school education is .356 (=353/993). Similarly, the proportion with a high school education is .508, and the proportion with a college education is .137. These are our weights. Then to get the adjusted, or standardized, percent militant among the very religious, we take the weighted sum of the percent militant across the three education groups that subdivide the "very religious" category (that is, the figures in the top row of the table): 17%*.356

TABLE 2.3. Percent Militant by Religiosity, and Percent Militant by Religiosity Adjusting (Standardizing) for Religiosity Differences in Educational Attainment, Urban Negroes in the U.S., 1964 (N = 993).

	Percent Militant	Percent Militant Adjusted for Education	N
Very religious	27	29	(230)
Somewhat religious	30	31	(532)
Not very or not at all religious	48	45	(231)
Percentage spread	21	16	

+ 34%*.508 + 38%*.137 = 29%. To get the adjusted percent militant among the "somewhat religious," we apply the same weights to the percentages in the second row in the table: 22%*.356 + 32%*.508 + 48%*.137 = 31%. Finally, to get the adjusted percent militant among the "not very or not at all religious," we do the same for the third row of the table, which yields 45 percent. We can then compare these percentages to the corresponding percentages for the zero-order relationship between religiosity and militancy (that is, not controlling for education). The comparison is shown in Table 2.3. (The Stata `-do-` file used to carry out the computations, using the command `-dstdize-` and the `-log-` file that shows the results, are available as downloadable files from the publisher, Jossey-Bass/Wiley (www.josseybass.com/go/quantitativedataanalysis) as are similar files for the remaining worked examples in the chapter. Because we have not yet begun computing, it probably is best to note the availability of this material and return to it later unless you are already familiar with Stata.)

STATA `-DO-` FILES AND `-LOG-` FILES In Stata, `-do-` files are commands, and `-log-` files record the results of executing `-do-` files. As you will see in Chapter Four, the management of data analysis is complex and is much facilitated by the creation of `-do-` files, which are efficient and also provide a permanent record of what you have done to produce each tabulation or coefficient. Anyone who has tried to replicate an analysis performed several years or even several months earlier will appreciate the value of having an exact record of the computations used to generate each result.

When presenting data of this sort, it is sometimes useful to compare the range in the percentage positive (in this case, the percent militant) across categories of the independent variable, with and without controls. In Table 2.3, we can see that the difference in the percent militant between the least and most religious categories is twenty-one points whereas, when education is controlled, the difference is reduced to sixteen points, a 24 percent reduction (= 1 − 16/21). In some sense, then, we can say that education "explains" about a quarter of the relationship between religiosity and militancy. We need to be cautious about making computations of this sort and only employ them when they are helpful in making the analysis clear. For example, it doesn't make much sense to compute a "spread" or "range" in the percentages if the relationship between religiosity and militancy is not *monotonic* (that is, if the percentage militant does not increase, or at least not decrease, as religiosity declines).

DIRECT STANDARDIZATION IN EARLIER SURVEY RESEARCH Although direct standardization is a conventional technique in demographic analysis, it also appears in the early survey research literature as a vehicle for getting to a "weighted net percentage difference" or "weighted net percentage spread." The really useful part of the procedure is the computation of adjusted, or standardized, rates. The subsequent computation of percentage differences or percentage spreads is only sometimes useful, as a way of summarizing the effect of control variables.

Example 2: Belief That Humans Evolved from Animals (Direct Standardization with Two or More Control Variables)

Sometimes we want to adjust, or standardize, our data by more than one control variable at a time to get a summary of the effect of some variable on another when two or more other variables are held constant. Consider, for example, acceptance of the scientific theory of evolution. In 1993, 1994, and 2000, the NORC *General Social Survey* (GSS) included a question:

> *For each statement below, just check the box that comes closest to your opinion of how true it is. In your opinion, how true is this? . . . Human beings developed from earlier species of animals.*

Table 2.4 shows the distribution of responses for the three years combined. Because there was virtually no difference between years, data from all three years have been combined to increase the sample size. Procedures for assessing variation across years are discussed in Chapter Seven. A description of the GSS and details on how to obtain GSS data sets are given in Appendix A. The Stata `-do-` file used to create this example and the resulting `-log-` file can be downloaded from the publisher's Web site, noted earlier. In general, the `-do-` and `-log-` files for each worked example are available for downloading.

To me, Table 2.4 is startling. How can it be that the overwhelming majority of American adults fail to accept something that is undisputed in the scientific community?

TABLE 2.4. **Percentage Distribution of Beliefs Regarding the Scientific View of Evolution (U.S. Adults, 1993, 1994, and 2000).**

Evolutionary explanation is . . .	
Definitely true	15.4%
Probably true	32.3
Probably not true	16.8
Definitely not true	35.5
Total	**100.0%**
N	(3,663)

Perhaps the recent increase in the proportion of the population that adheres to fundamentalist religious beliefs, especially fundamentalist Protestant views in which the Bible is taken as literally true, accounts for this outcome. To see whether this is so, in Table 2.5 I cross-tabulated acceptance of the scientific view of evolution (measured by endorsement of the statement that descent from other animals is "definitely true") by religious denomination, making a distinction between "fundamentalist" and "denominational" Protestants. (For want of better information, I simply dichotomized Protestant denominations on the basis of the proportion of their members in the sample who believe that "The Bible is the actual word of God and is to be taken literally, word for word." Denominations for which at least 50 percent of respondents gave this response—"other" Protestants and all Baptists except members of the "American Baptist Church in the U.S.A." and "Baptists, don't know which"—were coded as fundamentalist; all other Protestant denominations were coded as Denominational Protestants.) Unfortunately, the same distinction, between religious subgroups with and without a literal belief in the holy scriptures of their faith, cannot be made for non-Protestants given the way the data were originally coded in the GSS.

Although there are substantial differences among religious groups in their acceptance of the scientific view of evolution, the fundamentalist-denominational split among Protestants does not seem to be central to the explanation because there is only a 4 percent difference between the two groups. Interestingly, non-Christians appear to be much more willing than Christians to accept an evolutionary perspective, and Catholics appear to be more willing than Protestants to do so.

TABLE 2.5. **Percentage Accepting the Scientific View of Evolution by Religious Denomination (N = 3,663).**

	Percentage Accepting the Evolution of Humans from Animals as "Definitely True"	N
Fundamentalist Protestants	8.0	(968)
Denominational Protestants	11.8	(1,222)
Catholics	17.8	(858)
Other Christians	5.6	(18)
Jews	38.6	(83)
Other religion	23.6	(123)
No religion	32.5	(391)

Given these patterns, it could well be that the observed religious differences are, at least in part, spurious. In particular, educational differences among religious groups—Jews are particularly well educated and fundamentalist Protestants are particularly poorly educated—might partly account for religious differences in acceptance of the scientific view. Similarly, age differences among religious groups—the young are particularly likely to reject religion—might provide part of the explanation as well.

To consider these possibilities, we need to determine, first, whether acceptance of the scientific explanation of human evolution varies by age and education and, if so, whether religious groups differ with respect to their age and education. Tables 2.6 and 2.7 provide the necessary information regarding the first question, and Tables 2.8 and 2.9 provide the corresponding information regarding the second question.

Unsurprisingly, endorsement of the statement that humans evolved from other animals as "definitely true" increases sharply with education, as we see in Table 2.6, ranging from 9 percent of those with no more than a high school education to 36 percent of those with post-graduate education. It is also true that younger people are more likely to endorse the scientific explanation of evolution than are older respondents (see Table 2.7): 18 percent of those under age fifty, compared to 7 percent of those seventy and over, say that it is "definitely true" that humans evolved from other animals.

As expected, Table 2.8 shows that Jews are by far the best-educated religious group, followed by other non-Christian groups, and that Fundamentalist Protestants and Other Christians are the least well educated. Also as expected, Table 2.9 shows that those without religion tend to be young. However, members of "other" religious groups also tend—disproportionately—to be young, perhaps because they are mainly immigrants.

TABLE 2.6. **Percentage Accepting the Scientific View of Evolution by Level of Education.**

	Percentage	N
High school or less	9.2	(1,743)
Some college	11.9	(936)
College graduate	24.8	(561)
Post-graduate education	36.2	(423)

TABLE 2.7. **Percentage Accepting the Scientific View of Evolution by Age.**

	Percentage	N
18–49	17.5	(2,373)
50–69	13.5	(889)
70+	7.0	(401)

These results suggest that differences among religious groups with respect to age and education might, indeed, explain part of the observed difference in acceptance of the scientific view of evolution.

To see to what extent age and educational differences among religious groups account for religious group differences in acceptance of evolution, we can directly standardize the religion/evolution-beliefs relationship for education and age. We do this by determining the joint distribution of the entire sample with respect to age and education and then use the proportion in each age-by-education category as weights with which to compute, separately for each religious group, the weighted average of the age-by-education-specific percentages accepting a scientific view of evolution. By doing this, we treat each religious group as if it had exactly the same joint distribution with respect to age and education as did the entire sample. This procedure thus adjusts the percentage of each religious group that endorses the scientific perspective on evolution to remove the effect of religious group differences in the joint distribution of age and education.

TABLE 2.8. **Percentage Distribution of Educational Attainment by Religion.**

	High School or Less	Some College	College Graduate	Post-Graduate	Total	N
Fundamentalist Protestants	55.7	26.4	10.7	7.1	99.9	(968)
Denominational Protestants	47.6	26.5	15.2	10.6	99.9	(1,222)
Catholics	45.6	24.1	18.1	12.2	100.0	(858)
Other Christians	61.1	16.7	16.7	5.6	100.0	(18)
Jews	15.7	21.7	31.3	31.3	100.0	(83)
Other religion	36.6	25.2	19.5	18.7	100.0	(123)
No religion	41.4	24.8	16.1	16.6	99.9	(391)
Total	**47.6**	**25.6**	**15.3**	**11.6**	**100.1**	**(3,663)**

TABLE 2.9. **Percentage Distribution of Age by Religion.**

	18–49	50–69	70+	Total	N
Fundamentalist Protestants	59.6	27.9	12.5	100.0	(968)
Denominational Protestants	59.8	25.0	15.1	99.9	(1,222)
Catholics	67.8	24.8	7.3	99.9	(858)
Other Christians	88.9	11.1	0.0	99.9	(18)
Jews	61.4	25.3	13.2	99.9	(83)
Other religion	83.7	13.8	2.4	99.9	(123)
No religion	80.0	15.4	4.6	100.0	(391)
Total	**64.8**	**24.3**	**11.0**	**100.1**	**(3,663)**

To get the necessary weights, we simply cross-tabulate age by education and express the number of people in each cell of the table as a proportion of the total. These proportions are shown in Table 2.10.

We then tabulate the percentage accepting the scientific position on evolution by religion, age, and education. These percentages are shown in Table 2.11. Note that many of these percentages are based on very few cases. This means that they are not very precise, in the sense that they are subject to large sampling variability. We could collapse the education and age categories still further, but that would ignore substantial within-category heterogeneity. As always, there is a balance to be struck between sampling precision and substantive sensibility or—in terminology we will adopt later—between *reliability* and *validity*. In the present case, I might have been better advised to take a more conservative approach, especially because the cell-specific percentages bounce around a lot (exactly as we would expect given the large degree of sampling variability), which makes the differences in the resulting standardized percentages somewhat less clear cut. On the other hand, the weights are very small for the cells based on few cases, which minimizes their contribution to the overall percentages.

Finally, to get the adjusted, or standardized, coefficients, we sum the weighted percentages, where the weights are the proportions in Table 2.10. For example, the adjusted, or directly standardized, percentage of Fundamentalist Protestants who accept the evolutionary viewpoint as "definitively true" is, within rounding error,

$$\begin{aligned} 9.7 = {} & 5.7*.274 + 3.8*.184 + 15.7*.110 + 29.3*.080 \\ & + 4.9*.126 + 3.3*.056 + 25.0*.032 + 40.9*.029 \\ & + 3.3*.076 + 7.7*.015 + 10.0*.012 + 16.7*.007 \end{aligned}$$

The remaining standardized percentages are derived in the same way. They are shown in Table 2.12, with the observed percentages repeated from Table 2.5 to make comparisons easier.

TABLE 2.10. **Joint Probability Distribution of Education and Age.**

	18–49	50–69	70+	Total
High school or less	.274	.126	.076	**.476**
Some college	.184	.056	.015	**.256**
College graduate	.110	.032	.012	**.153**
Post-graduate education	.080	.029	.007	**.116**
Total	**.648**	**.243**	**.110**	**1.001**

TABLE 2.11. **Percentage Accepting the Scientific View of Evolution by Religion, Age, and Sex (Percentage Bases in Parentheses).**

	Fundamentalist Protestants	Denominational Protestants	Catholic	Other Christians	Jewish	Other	None
Age 18–49							
High school or less	5.7 (283)	11.8 (321)	9.1 (220)	[10.0] (10)	[25.0] (4)	21.1 (38)	18.6 (129)
Some college	3.8 (183)	10.1 (208)	20.1 (159)	[0.0] (2)	[28.6] (14)	11.5 (26)	29.3 (82)
College graduate	15.7 (70)	19.3 (119)	30.4 (125)	[0.0] (3)	[41.2] (17)	35.0 (20)	40.4 (47)
Post-graduate education	29.3 (41)	25.3 (83)	37.2 (78)	[0.0] (1)	[62.5] (16)	[36.8] (19)	58.2 (55)
Age 50–69							
High school or less	4.9 (164)	6.2 (146)	11.8 (119)	[0.0] (1)	[75.0] (4)	[0.0] (4)	22.7 (22)
Some college	3.3 (60)	9.0 (78)	6.8 (44)	[0.0] (1)	[25.0] (4)	20.0 (6)	[21.4] (14)
College graduate	25.0 (24)	21.3 (47)	20.0 (25)	– (0)	[20.0] (5)	[25.0] (4)	[41.7] (12)
Post-graduate education	40.9 (22)	20.0 (35)	32.0 (25)	– (0)	[37.5] (8)	[25.0] (4)	[66.7] (12)

(Continued)

TABLE 2.11. **Percentage Accepting the Scientific View of Evolution by Religion, Age, and Sex (Percentage Bases in Parentheses).** *(Continued)*

	Fundamentalist Protestants	Denominational Protestants	Catholic	Other Christians	Jewish	Other	None
Age 70 or more							
High school or less	3.3 (92)	1.7 (115)	5.8 (52)	– (0)	[20.0] (5)	[0.0] (3)	[27.3] (11)
Some college	[7.7] (13)	5.3 (38)	[0.0] (4)	– (0)	– (0)	– (0)	[0.0] (1)
College graduate	[10.0] (10)	10.0 (20)	[20.0] (5)	– (0)	[0.0] (4)	– (0)	[50.0] (4)
Post-graduate education	[16.7] (6)	[16.7] (12)	[0.0] (2)	– (0)	[50.0] (2)	– (0)	[100.0] (2)

Note: Percentages based on fewer than twenty cases (shown in square brackets) should be interpreted with caution.

TABLE 2.12. **Observed Proportion Accepting the Scientific View of Evolution, and Proportion Standardized for Education and Age.**

	Percentage Accepting Scientific View of Evolution as "Definitely True"	Percentage Accepting Scientific View, Standardized by Age and Education	N
Fundamentalist Protestants	8.0	9.7	(968)
Denominational Protestants	11.8	12.2	(1,222)
Catholics	17.8	16.6	(858)
Other Christians	5.6	2.7	(18)
Jews	38.6	36.0	(83)
Other religion	23.6	19.9	(123)
No religion	32.5	30.2	(391)

As you can see, despite the association between religious group affiliation and, respectively, age and education, and the association of age and education with acceptance of an evolutionary account of human origins, standardizing for these variables has relatively little impact on religious group differences in acceptance of the claim that humans evolved from other animals. The one exception is the non-religious, whose support for a scientific view of evolution appears to be due, in part, to their relatively young age. Despite minor shifts in the expected direction for Fundamentalist Protestants and for Jews and those with other religions, the dominant pattern is one of religious group differences in acceptance of an evolutionary view of the origins of mankind that are *not* a simple reflection of religious differences in age and education but presumably reflect, instead, the theological differences that distinguish religious categories.

Example 3: Occupational Status by Race in South Africa

Now let us consider another example: the extent to which racial differences in occupational attainment in South Africa can be explained by racial differences in education (the data are from the *Survey of Economic Opportunity and Achievement in South Africa*, conducted in the early 1990s [Treiman, Lewin, and Lu 2006]; the Stata -do- and -log- files for the worked example are available as downloadable files; for information on the data set and how to obtain it, see Appendix A). From the left-hand panel of Table 2.13, it is evident that there are strong differences in occupational attainment by race. Non-Whites, especially Blacks, are substantially less likely to be managerial, professional, or technical workers than are Whites and are substantially more likely to be semiskilled or unskilled manual workers. Moreover, Blacks are far more likely to be unemployed than are members of any other group. It is also well known that substantial racial differences exist in educational attainment in South Africa, with Whites by far the best educated, followed, in order, by Asians (who in South Africa are mainly descendants of people brought as indentured workers from the Indian subcontinent), Coloureds (mixed-race persons), and Blacks (these are the racial categories conventionally used in South Africa); and also that in South Africa, as elsewhere, occupational attainment depends to a considerable degree on educational attainment (Treiman, McKeever, and Fodor 1996). Under these circumstances, we might suspect that racial differences in occupational attainment can be largely explained by racial differences in educational attainment. Indeed, this is what Treiman, McKeever, and Fodor (1996) found using the International Socioeconomic Status Index (ISEI) (Ganzeboom, de Graaf, and Treiman 1992; Ganzeboom and Treiman 1996) as an index of occupational attainment. However, it also is possible that access to certain types of occupations, such as professional and technical positions, depends heavily upon education, whereas access to others, such as managerial positions, may be denied on the basis of race to those who are educationally qualified.

To determine to what extent, and for which occupation categories, racial differences in access can be explained by racial differences in education, I adjusted (directly standardized) the relationship between race and occupational status by education. Here I used the White distribution of education, computed from the weighted data, as the standard distribution to determine what occupational distributions for each of the non-White groups might be expected were they able to upgrade their levels of educational attainment so that they had the same distributions across schooling levels as did Whites.

The results are shown in panel B of Table 2.13. They are quite instructive. Bringing the other racial groups to the White distribution of education (and assuming that doing so would not affect the relationship between education and occupational attainment within each group), racial differences in the likelihood of being a professional would entirely disappear. Indeed, Blacks would be slightly more likely than members of the other groups to become professionals. By contrast, the percentage of each race group in the managerial category would remain essentially unchanged, suggesting that it is not education but rather norms about who is permitted to supervise whom that account for the racial disparity in this category. The remaining large changes apply to only one or two of the three non-White groups: Asians would not be very substantially affected except for a reduction in the proportion semiskilled; Coloureds would increase the proportion in technical jobs and reduce the proportion in semiskilled and unskilled manual jobs and farm labor; and Blacks would increase the proportion in clerical jobs and reduce the proportion in all manual categories. If all four racial groups had the same educational distribution as Whites, the dissimilarity (measured by Δ; see Chapter Three) between the occupational distributions of Whites and Asians would be reduced by about 30 percent (from 29.2 to 20.5) as would the dissimilarity in the occupational distributions of Whites and Coloureds (from 37.9 to 26.5), whereas the dissimilarity in the occupational distributions of Whites

TABLE 2.13. Percentage Distribution of Occupational Groups by Race, South African Males Age 20–69, Early 1990s (Percentages Shown Without Controls and also Directly Standardized for Racial Differences in Educational Attainment;[a] N = 4,004).

	Without Controls				Adjusted for Education			
	White	Asian	Coloured	Black	White	Asian	Coloured	Black
Managers	18.4	9.7	3.3	0.7	18.1	10.5	5.2	1.5
Professionals	13.7	7.0	5.3	3.3	13.2	13.6	11.9	16.5
Technical	17.1	11.6	5.0	2.0	16.9	13.3	8.5	2.4
Clerical	7.2	13.4	7.1	5.8	7.2	13.0	9.1	8.5
Service	4.0	3.9	4.0	6.8	4.0	2.8	4.3	6.7
Sales	2.2	8.2	2.8	3.5	2.2	9.4	2.5	2.6
Farm	4.6	0.5	2.3	0.8	4.5	0.3	1.1	0.3
Skilled manual	18.5	21.1	22.2	14.4	18.6	18.8	19.3	8.2
Semiskilled	3.2	12.9	10.4	16.7	3.2	7.6	6.6	9.9
Unskilled manual	0.6	3.6	14.4	13.5	0.6	2.6	7.6	6.2
Farm laborers	0.0	0.2	7.8	2.3	0.0	0.0	3.1	0.4
Occupation unknown[b]	9.8	5.6	11.4	13.0	9.7	6.2	16.6	20.1
Unemployed	0.9	2.2	4.0	17.2	0.9	1.8	4.4	16.6
Total	**100.2**	**99.9**	**100.0**	**100.0**	**99.1**	**99.9**	**100.0**	**100.0**
N	(1236)	(412)	(396)	(1960)	(1236)	(412)	(396)	(1960)
Dissimilarity from Whites (Δ)[c]	29.2	37.9	52.4		20.5	26.5	46.1	

[a] The standard population is the educational distribution of the White male population—computed from the survey data weighted to conform to the census distributions of region by urban versus rural residence.

[b] Because coding of occupation data to detailed occupations had not been completed when this table was prepared, I have included a separate category "occupation unknown."

[c] Index of Dissimilarity = 1/2 the sum of the absolute values of the differences between the percentage of Whites and the percentage of the other racial group in each occupation category. See Chapter Three for further exposition of this index.

and Blacks would be reduced only about 12 percent (from 52.4 to 46.1). The substantial remaining dissimilarity in the occupational distributions of the four race groups net of education suggests that Treiman, McKeever, and Fodor's (1996) conclusion that education largely explains occupational *status* differences between race groups in South Africa does not tell the whole story.

Example 4: Level of Literacy by Urban Versus Rural Residence in China

Now consider a final example—the relationship among education, urban residence, and degree of literacy in the People's Republic of China. In a 1996 national sample of the adult population (Treiman, Walder, and Li 1996), respondents were asked to identify ten Chinese characters (see Appendix A regarding the properties of this data set and how to obtain access to it). The number of correct identifications is interpreted as indicating the degree of literacy (Treiman 2007a). Obviously, literacy would be expected to increase with education. Moreover, I would expect the urban population to score better on the character recognition task just because urban respondents tend to get more schooling than do rural respondents. The question of interest here is whether educational differences between the rural and urban population entirely explain the observed mean difference in the number of characters correctly identified, which is 1.8 (as shown in Table 2.14). To determine this, I adjusted (directly standardized) the urban and rural means by assuming that both populations have the same distribution of education—the distribution for the entire adult population of China, computed from the weighted data. Note that in this example it is not percentages that are standardized but rather means. The procedure is identical in both cases, although if this is done by computer (using Stata), a special adjustment needs to be made to the data to overcome a limitation in the Stata command—the requirement that the numerators of the "rates" to be standardized (what Stata calls `-charvar-`) be integers. To see how to do this, consult the Stata `-do-` and `-log-` files for this example, which are included in the set of downloadable files for this chapter.

TABLE 2.14. Mean Number of Chinese Characters Known (Out of 10), for Urban and Rural Residents Age 20–69, China 1996 (Means Shown Without Controls and Also Directly Standardized for Urban-Rural Differences in the Distribution of Education;[a] N = 6,081).

	Without Controls	Adjusted for Education	N
Urban residents	4.8	4.0	(3,079)
Rural residents	2.0	2.4	(3,002)
Difference	2.8	1.6	

[a] The standard population is the entire population of China age 20–69, computed from the survey data weighted to reflect differential sampling rates for the rural and urban populations and to correct for variations in household size. Nine cases for which data on education were missing were omitted.

The results are quite straightforward and require little comment. When education is standardized, the urban-rural gap in the mean number of characters correctly identified is reduced from 2.8 to 1.6. Thus, about 43 percent (= 1 − 1.6/2.8) of the urban-rural difference in vocabulary knowledge is explained by rural-urban differences in the level of educational attainment.

Although the four examples presented here all standardize for education, this is purely coincidental. Many other uses of direct standardization are imaginable. For example, it probably would be possible to explain higher crime rates among early twentieth-century immigrants to the United States than among natives simply by standardizing for age and sex. Immigrants were disproportionately young males, and young males are known to have higher crime rates than any other age-sex combination.

A FINAL NOTE ON STATISTICAL CONTROLS VERSUS EXPERIMENTS

In describing the logic of cross-tabulations, I have been describing the logic of nonexperimental data analysis in general. True experiments are relatively uncommon in social research, although they are widely used in psychological research and increasingly in microeconomics (for a very nice example of the latter, see Thomas and others [2004]). A true experiment is a situation in which the objects of the experiment are randomly divided into two or more groups, and one group is exposed to some treatment while the other group is not, or several groups are exposed to different treatments. If the groups then differ in some outcome variable, the differences can be attributed to the differences in treatments. In such cases we can unambiguously establish that the treatment caused the difference in outcomes (although we may not know the exact mechanism involved). (Of course, this claim holds only when differences between the experimental and control groups are not inadvertently introduced by the investigators as a consequence of design flaws or of failure to rigorously adhere to the randomized trial design. For a classic discussion of such problems, see Campbell and Stanley [1966] or a shorter version by Campbell [1957] that contains the core of the Campbell and Stanley material.)

When experiments are undertaken in fields such as chemistry, sampling is not ordinarily a consideration because it can be confidently assumed that any batch of a chemical will behave like any other batch of the same chemical; only when things go wrong do chemists tend to question that assumption. In the social and behavioral and many of the biological sciences, by contrast, it cannot be assumed that one subject is just like another subject. Hence, in experiments in these fields, subjects are randomly assigned to treatment groups. In this way, it becomes possible to assess whether group differences in outcomes are larger than would be likely to occur by chance because of sampling variability. If so, we can say, subject only to the uncertainty of statistical inference, that the difference in treatments caused the difference in outcomes.

In the social sciences, random assignment of subjects to treatment groups is often—in fact usually—impossible for several reasons. First, both ethical and practical considerations limit the kind of experimentation that can be done on human subjects. For example, it would be neither ethical nor practically possible to determine whether one sort of schooling was pedagogically superior to another by randomly assigning children to different schools and several years later determining their level of educational achievement. In addition, many phenomena of interest to social scientists are simply not experimentally manipulable, even in principle. The propensity for in-group solidarity to increase in wartime, for example, is not something that can be experimentally confirmed, nor can the proposition that social stratification is more pronounced in sedentary agricultural societies than in hunter-gatherer societies.

Occasionally, "natural experiments" can be analyzed. Natural experiments are situations in which different individuals are exposed to different circumstances, and it can be reasonably assumed that the circumstance to which individuals are exposed is essentially random. A very nice example of such an analysis is the test by Almond (2006) of the "fetal origins hypothesis." He showed convincingly that individuals in utero during the few months in which the 1918 flu pandemic was raging suffered reduced educational attainment, increased rates of physical disability, and lower income in midlife relative to those in utero in the few months preceding and following the epidemic. Because there is no basis for expecting the exact month of conception to be correlated with vulnerability to the flu virus, the conditions of a natural experiment were well satisfied in this elegant analysis. Natural experiments have become increasingly popular in economics as the limitations of various statistical fixes to correct for "sample selection bias" have become more evident. We will return to this issue in the final chapter. (For additional examples of natural experiments that are well worth reading, see Campbell and Ross [1968], Berelson [1979], Sloan and others [1988], and the papers cited in Chapter Sixteen.)

Given the limited possibilities for experimentation in the social sciences, we resort to a variety of statistical controls of the sort discussed here and later. These procedures share a common logic: they are all designed to hold constant some variable or variables so that the net effect of a given variable on a given outcome can be assessed.

THE WEAKNESS OF MATCHING AND A USEFUL FIX

Sometimes survey analysts attempt to simulate random assignment by matching comparison groups on some set of variables. In its original form, this practice was inherently unsatisfactory. When attempting to match on all potentially relevant factors, it is difficult to avoid running out of cases. Moreover, no matter how many variables are used in the match, it is always possible that the experimental and control groups differ on some nonmatched factor that is correlated with the experimental outcome. However, combining matching with statistical controls can be a useful strategy, especially when the adequacy of the match is summarized via a "propensity score" (Rosenbaum and Rubin 1983). For recent treatments of propensity score matching, see Smith (1997), Becker and Ichino (2002), Abadie and others (2004), Brand (2006), Brand and Halaby (2006), and Becker and Caliendo (2007). Harding (2002) is an instructive application. Propensity score matching is also discussed in Chapter Sixteen.

Compared to experiments, statistical controls have two fundamental limitations, which make it impossible to definitively prove any causal theory (although definitive *disproof* is possible). First, no matter how many control variables we introduce, we can never be sure that the remaining net relationship is a true causal relationship and not the spurious result of some yet-to-be-introduced variable.

Second, although we speak of *holding constant* some variable, or set of variables, what we usually do in practice is simply reduce the within-group variability for these variables. This is particularly obvious when we are dealing with cross-tabulations because we generally divide the sample into a small set of categories. In what sense, for example, can we be said to "hold education constant" when our categories consist of those with less than a high school education, those with some sort of high school experience, and those with some sort of college experience? Although the within-category variability in educational attainment obviously is smaller than the total variability in the sample as a whole, it is still substantial. Hence, if two other variables both depend on educational attainment, they are likely to be correlated within educational categories as gross as these, as well as across educational categories. As you will see in more detail later, using interval or ratio variables in a regression framework will not solve the problem but merely transform it. Although the within-category variability generally will be reduced, the very parsimony in the expression of relationships between variables that regression procedures permit will generally result in some distortion of the true complexities of such relationships—discontinuities, nonlinearities, and so on, only some of which can be represented succinctly.

Our only salvation is adequate theory. Because we can seldom definitively establish causal relationships by reference to data, we need to build up a body of theory that consists of a set of plausible, mutually consistent, empirically verified propositions. Although we cannot definitively prove causal relations, we can determine whether our data are *consistent with* our theories; if so, we can say that the proposition is tentatively empirically verified. We are in a stronger position when it comes to *disproof.* If our data are *inconsistent* with our theory, that usually is sufficient grounds for rejecting the theory, although we need to be sensitive to the possibility that there are omitted variables that would change our conclusions if they were included in our cross-tabulation or model. In short, to maintain a theory, it is *necessary* but not *sufficient* that the data be as predicted by the theory. Because consistency is necessary to maintain the theory, inconsistency is sufficient to require us to reject it—provided we can be confident that we have not omitted important variables. (On the other hand, as Alfred North Whitehead is supposed to have said, never let data stand in the way of a good theory. If the theory is sufficiently strong, you might want to question the data. I will have more to say about this later, in a discussion of concepts and indicators.)

WHAT THIS CHAPTER HAS SHOWN

In this chapter we have considered the logic of multivariate statistical analysis and its application to cross-tabulations involving three or more variables. The notion of an interaction effect—a situation in which the effect of one independent variable depends on the value or level of one or more other independent variables—was introduced. This is a very important idea in statistical analysis, and so you should be sure that you understand it thoroughly. We also considered suppressor effects, situations in which the effect of one independent variable offsets the effect of another independent variable because the two effects have opposite signs. In such situations, the failure to include both variables in the model can lead to an understatement of the true relationships between the included variable and the dependent variable. We then turned to direct standardization (sometimes called covariate adjustment), a procedure for purging a relationship of the effect of a particular variable or variables. Direct standardization can be thought of as a procedure for creating "counterfactual" or "what if" relationships—for example, what would be the relationship between religiosity and militancy if we adjusted for the fact that well educated Blacks in the 1960s tended to be both more religious and less militant than less well educated Blacks. Having discussed the logic of direct standardization, we considered several technical aspects of the procedure to see how to standardize data starting not only from tables but also from individual records; to standardize percentage distributions; and to standardize means. We concluded by considering the limitations of statistical controls, in contrast to randomized experiments.

In the following chapter, we complete our initial discussion of cross-tabulation tables by considering how to extract new information from published tables; then note the one circumstance in which it makes sense to percentage a table "backwards"; touch for the first but not for the last time on how to handle missing data; consider cross-tabulation tables in which the cell entries are means; present a measure of the similarity of percentage distributions, the Index of Dissimilarity (Δ); and end with some comments about how to write about cross-tabulations.

CHAPTER 3

STILL MORE ON TABLES

WHAT THIS CHAPTER IS ABOUT

In this chapter we wrap up our discussion of cross-tabulations for now. After spending some time learning to love the computer—a very brief time, actually—and then delving into the mysteries of regression equations, we will return to cross-tabulations and discuss procedures for making inferences about relations embodied in them via log-linear analysis.

We begin this chapter with a discussion of how to extract new information from published tables; then note the one circumstance in which it makes sense to percentage a table "backwards"; touch for the first but not for the last time on how to handle missing data; consider cross-tabulation tables in which the cell entries are means; present a measure of the similarity of percentage distributions, the Index of Dissimilarity (Δ); and end with some comments about how to write about cross-tabulations.