

# 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.



## 4 Quantitative Data Analysis: Doing Social Research to Test Ideas

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 a 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         | <b>230</b> |
| Somewhat religious   | 160        | 372         | <b>532</b> |
| Not very religious   | 87         | 108         | <b>195</b> |
| Not at all religious | 25         | 11          | <b>36</b>  |
| <b>Total</b>         | <b>333</b> | <b>660</b>  | <b>993</b> |

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%                  | <b>33%</b>  |
| Nonmilitant  | 73             | 70                 | 55                 | 31                   | <b>67</b>   |
| <b>Total</b> | <b>100%</b>    | <b>100%</b>        | <b>100%</b>        | <b>100%</b>          | <b>100%</b> |
| 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 ( $\times 100$ ) 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.**

| Religiosity          | Educational Attainment |             |             |
|----------------------|------------------------|-------------|-------------|
|                      | 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          |
| <b>Total</b>         | <b>101%</b>            | <b>101%</b> | <b>100%</b> |
| 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.**

| Militancy    | Educational Attainment |             |             |
|--------------|------------------------|-------------|-------------|
|              | Grammar School         | High School | College     |
| Militant     | 22%                    | 36%         | 53%         |
| Nonmilitant  | 78                     | 64          | 47          |
| <b>Total</b> | <b>100%</b>            | <b>100%</b> | <b>100%</b> |
| 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.**

| Militancy    | Grammar School |             |             | High School |             |             | College     |             |             |
|--------------|----------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
|              | 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          |
| <b>Total</b> | <b>100%</b>    | <b>100%</b> | <b>100%</b> | <b>100%</b> | <b>100%</b> | <b>100%</b> | <b>100%</b> | <b>100%</b> | <b>100%</b> |
| 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).**

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

