CHAPTER

15

Construction and Analysis of Contingency Tables

Chapter 2, on the topic of measurement, introduced the three levels of measurement—nominal, ordinal, and interval—and discussed the measures of central tendency that can be used to describe and summarize variables of each type. Although this information provides a useful guide to the treatment of single variables, ordinarily such univariate statistics constitute only the first step in data analysis—and in the job of the public or nonprofit manager.

For example, imagine for a moment that you work in the department of public affairs for a large government or nonprofit agency. The department has just finished conducting its annual survey of public opinion about the agency. Some of the initial results show that most of the people interviewed now feel that the agency is doing a "very poor job," and the median opinion is not very cheery, either—a "poor job." This assessment represents a dramatic downturn in public opinion compared with previous years. To be sure, this is important information, but obviously it is not the kind of news that you would want to give to your boss or to the mayor and the budget-minded city council or board of directors *without some idea of how the public image of the agency might be improved*. But how might this goal be attained?

One way to approach this question is to consider *why* public support has fallen. There may be several reasons. Perhaps the agency has cut a popular program that it used to administer in Avery County, California, one of several counties over which it has jurisdiction. If the loss of this program is responsible for the drop in agency prestige, then you would expect to find a lower level of public favor in Avery County than in other counties where it has not been necessary to cut programs. Or perhaps the fall in public esteem is a result of the recent appointment of a new director of the agency, "mean" Gene Medford, whose past political exploits received rough treatment in the local press. If so, then you would hypothesize that those citizens who disapproved of the appointment would be more critical of the job performance of the agency than would those who approved of the appointment. Fortunately, the survey of public opinion conducted by the agency elicited information pertaining to citizen residence and attitude toward the new director, so both of these ideas can be checked out.

These proposed explanations for the decline in public opinion carry different implications for public policy. If data analysis yielded support for the first

explanation, then the chief executive could be informed (gently!) that although public opinion of the agency is low, evidence indicates that it could be improved through restoration of the program that had been cut in Avery County. On the other hand, if the data showed support for the second explanation, the chief executive might advise Medford to clear the air about his past through public speeches and press conferences—or the chief executive might decide that less pleasant steps are necessary.

Regardless of which (if either) explanation turns out to be correct, the important point to bear in mind is that data analysis has moved from a concern with a single variable—public opinion toward the performance of the agency—to a focus on *relationships between variables*. This sequence is typical in the analysis of data. Generally, we would like to know not only the distribution of scores or responses on a variable of interest but also an explanation for this distribution. Is there a relationship between the county of residence of a citizen and attitude toward the agency? Is there a relationship between citizens' attitudes toward the new director and their attitudes toward the agency? In other words, do the responses on one variable help explain or account for responses on a second variable?

This chapter begins with the development of statistical methods to answer such questions. It is concerned with relationships between variables measured at the nominal and ordinal levels. (Relationships between interval-level variables are the subject of Part VI of this book.) The method that is generally employed to examine these relationships is called *contingency table analysis* or the analysis of *cross-tabulations*. In this chapter, we show how to set up a contingency table—cross-tabulating the responses to a pair of nominal or ordinal variables—and how to interpret it. The next two chapters elaborate on this topic: Chapter 16 presents aids to the interpretation of contingency tables, such as "measures of association" between variables. Chapter 17 discusses a procedure called *control table analysis*, or *statistical controls*, through which the relationships among three or more variables can be examined.

Percentage Distributions

Before we can treat the construction and interpretation of contingency tables, we need to review *percentage distributions* (see Chapter 4). A contingency table is actually a bivariate (two-variable) percentage (or frequency) distribution. If you do not need the review, congratulations—please skip to the next section.

A percentage distribution is simply a frequency distribution that has been converted to percentages. It tabulates the percentage associated with each data value or group of data values. Consider the distribution of responses of a sample of individuals to a standard survey question that asks respondents to consider whether there are too many bureaucrats in the federal government. The distribution is shown in Table 15.1.

As it stands, this table is difficult to interpret. Although it is evident that the mode is "agree" (the data value that occurs with greatest frequency; see Chapter 5

Table 15.1 Distribution of Responses

To what extent would you agree or disagree with the following statement? There are currently too many bureaucrats working for the federal government.

Response	Number of People
Strongly agree	686
Agree	979
Neutral	208
Disagree	436
Strongly disagree	232

on measures of central tendency), the table does not give a clear presentation of this opinion's popularity. Is it held by half of the people interviewed? A third? Nor does the table communicate the relative frequency of occurrence of the other opinions (strongly disagree, disagree, and so on). What proportion of the sample voiced these responses?

Without this information, it is difficult not only to comprehend this distribution of responses but also to compare it with other distributions of attitudes. For example, it would be interesting to know how this particular distribution of opinion toward federal bureaucrats compares with distributions obtained when the question was put to different samples of people, in different regions of the country, or at different times. Has there been a trend among different groups, or in different regions, or over time toward the view that there are too many federal bureaucrats? Does the public feel the same way about local bureaucrats or state bureaucrats?

The raw response figures displayed in Table 15.1 cannot answer these questions. In order to address them, data analysts conventionally convert the raw figures to percentages.

Steps in Percentaging

The procedure for converting raw figures to percentages involves three steps:

Step 1: Add the number of people (frequencies) giving each of the responses. In Table 15.1, this sum is equal to 686 + 979 + 208 + 436 + 232 = 2,541.

Step 2: Divide each of the individual frequencies by this total and multiply the result by 100. For example, for the response "strongly agree" in Table 15.1, we divide 686 by 2,541 and obtain .26997. Then we multiply this result by 100, yielding 26.997. This figure is the *percentage* of the people interviewed who gave the response "strongly agree." Repeat the procedure for each of the other response categories.

Table 15.2 **Percentage Distribution: Calculations**

To what extent would you agree or disagree with the following statement? There are currently too many bureaucrats working for the federal government.

Response	Frequency	Percentage	
Strongly agree	686	(686 ÷ 2,541) × 100 =	27.0
Agree	979	(979 ÷ 2,541) × 100 =	38.5
Neutral	208	(208 ÷ 2,541) × 100 =	8.2
Disagree	436	(436 ÷ 2,541) × 100 =	17.2
Strongly disagree	232	(232 ÷ 2,541) × 100 =	9.1
Total	2,541		100.0

Step 3: Round each of the percentages to one decimal place. If the second place to the right of the decimal point is greater than or equal to 5, add 1 to the first place to the right of the decimal. In this procedure, .16 becomes .2, .43 becomes .4, and 26.997 becomes 27.0. Table 15.2 shows the percentage distribution. (You may prefer to express the percentages as whole numbers, with no decimal places. Follow these same rules for rounding.)

Displaying and Interpreting Percentage Distributions

The percentage distribution displays the percentage of respondents giving each of the responses to the survey item. The only frequency or raw number that should be presented in the table is the total number of cases, usually abbreviated **N**. The total frequency helps the reader evaluate the distribution of responses. In general, the larger the number of cases on which the percentages are based, the greater the confidence in the results. For example, you would normally have more confidence in a percentage distribution based on 2,541 respondents than in one based on 541 or 41. Table 15.3 shows the final "percentaged" table.

The percentage distribution facilitates interpretation and comparison. It is clear from the percentage distribution in Table 15.3 that approximately 40% of those interviewed (the mode) "agree" that there are currently too many federal bureaucrats and that 65.5% (27.0% + 38.5% = 65.5%), or nearly two-thirds, express agreement with this notion (either "strongly agree" or "agree"). The extent of agreement far outweighs the extent of disagreement—65.5% versus 26.3% (the percentage indicating either "disagree" or "strongly disagree"; 17.2% + 9.1% = 26.3%). Only a small proportion (8.2%) remains "neutral."

These percentages can be compared with those obtained in other surveys of public opinion (for example, surveys conducted at other points in time or

Table 15.3 Percentage Distribution: Final Table

To what extent would you agree or disagree with the following statement? There are currently too many bureaucrats working for the federal government.

Response	Percentage
Strongly agree	27.0
Agree	38.5
Neutral	8.2
Disagree	17.2
Strongly disagree	9.1
Total	100.0
	(**N** = 2,541)

administered to different groups) to assess how attitudes toward bureaucrats are changing. For instance, if 5 years ago a similar survey of public opinion indicated that only 40% of the public expressed agreement that there are too many federal bureaucrats, it would be evident that public opinion is becoming more negative.

Collapsing Percentage Distributions

Often, public and nonprofit managers combine, or *collapse*, several of the original response categories in order to form a smaller number of new categories and to calculate percentages based on the new categories. For example, in the preceding discussion, the response categories "strongly agree" and "agree" and the categories "strongly disagree" and "disagree" were collapsed into broader categories of "agreement" and "disagreement," respectively.

To calculate percentages in a collapsed distribution, you employ the procedure described earlier: (1) Compute the total frequency, (2) divide the frequency of each of the new categories by this total and multiply by 100, and (3) round to the first decimal place. Alternatively, if the percentage distribution for the variable has already been computed based on the original response categories, the percentages for the new collapsed categories can be found by adding the percentages for the categories that have been collapsed. (The percentages for categories that have not been collapsed will not change.) The latter method was employed in the preceding discussion. For example, because 27.0% of the sample stated that they "strongly agree" that there are too many federal bureaucrats and 38.5% "agree," a total of 65.5% fall into the new collapsed category of "agree." The first of these methods for percentaging a collapsed distribution is illustrated in Table 15.4.

There are two primary reasons for presenting the percentage distribution in collapsed form. First, it is easier to interpret a distribution based on a few response categories than one based on many. In many instances, such as the preparation of memoranda, the collapsed distribution presents all the information managers

Table 15.4 Collapsed Percentage Distribution

To what extent would you agree or disagree with the following statement? There are currently too many bureaucrats working for the federal government.

Original Response Categories	(Original) Frequency	Collapsed Response Categories	(Collapsed) Frequency	Percentage	
Strongly agree	686	Agree	1,665	(1,665 ÷ 2,541) × 100 =	65.5
Agree	979				
Neutral	208	Neutral	208	(208 ÷ 2,541) × 100 =	8.2
Disagree	436				
Strongly disagree	232	Disagree	668	(668 ÷ 2,541) × 100 =	26.3
Total	2,541	Total	2,541		100.0

need to know, without burdening them with unnecessary complexity. Second, often in public and nonprofit administration, the data analyst or manager is not confident that the distinction between some response categories is very clear or meaningful; that is, you can generally be much more confident that, *in all*, 65.5% of those interviewed agree with a proposition than that *exactly* 27.0% "strongly agree" and *exactly* 38.5% "agree." In order to avoid communicating a false sense of precision, categories may be collapsed. Another good reason to get used to collapsed percentage distributions is that most contingency tables are based on this format (see below and Chapters 16 and 17).

When you collapse response categories of a variable, the collapsing must not pervert the meaning of the original categories. Response categories should be collapsed only if they are close in substantive meaning. Whereas the kind of collapsing we have done here—"strongly agree" and "agree", "strongly disagree" and "disagree"—is justified, collapsing the categories of "disagree" and "neutral" would not be.

The major exception to this rule occurs in distributions of *nominal* variables that have many response categories. Frequently only a few of the categories will have a large percentage of cases, whereas most of the categories will have only trivial numbers. In this situation, the analyst may choose to present each of the categories containing a substantial percentage and a category labeled "other," formed by collapsing all the remaining categories. For example, consider the variable "religion." In a given sample, the distribution of religion may be 62% Protestant, 22% Catholic, 13% Jewish, 1% Shinto, .5% Buddhist, .6% Hedonist, .5% Janist, and .4% Central Schwenkenfelter. To summarize this distribution, the analyst may present the percentages as shown in Table 15.5; note the use of a collapsed "other" category.

An exercise may be helpful to illustrate these points. The Shawnee Heights Independent Transit Authority has commissioned a poll of 120 persons to determine where Shawnee citizens do most of their shopping. This is important information in determining future transit routes in Shawnee Heights. The transit planners receive the data shown in Table 15.6.

Table 15.5 Collapsed Percentage Distribution for Religion

Religion	Percentage
Protestant	62
Catholic	22
Jewish	13
Other	3
Total	100
	(**N** = 1,872)

Table 15.6 Data for Shawnee Heights Poll

Main Store Named	Number of Persons
Cleo's (neighborhood store)	5
Morgan's (downtown)	18
Wiese's (eastern shopping center)	12
Cheatham's (neighborhood store)	2
Shop City (eastern shopping center)	19
Food-o-Rama (western shopping center)	15
Stermer's (downtown)	7
Binzer's (neighborhood store)	2
England's (western shopping center)	1
Bargainville's (eastern shopping center)	26
Whiskey River (downtown)	13
	120

In the space provided, construct a collapsed percentage distribution of the data in Table 15.6. (*Hint*: Consider collapsing categories based on common locations.)

Contingency Table Analysis

Analysis of contingency tables or cross-tabulations is the primary method researchers use to examine relationships between variables measured at the ordinal and nominal levels. The remainder of the chapter discusses the construction and interpretation of contingency tables. As you will see, the methods for percentaging are instrumental to this type of analysis.

Constructing Contingency Tables

A **contingency table** or **cross-tabulation** is a bivariate frequency distribution. We have dealt with **univariate**, or single-variable, frequency distributions in examples in this chapter and in previous chapters. A univariate frequency distribution presents the number of cases (or frequency) that has each value of a given variable. By analogy, a **bivariate**, or two-variable, frequency distribution presents the number of cases that fall into each possible pairing of the values or categories of two variables simultaneously. This definition is more readily visualized in an example.

Consider the cross-tabulation of the variables "race" (white, nonwhite) and "sex" (male, female) for volunteers to the Klondike Expressionist Art Museum. As these variables are defined here, there are four possible pairings: white and male, white and female, nonwhite and male, and nonwhite and female volunteers. Pairings across variables are easier to conceptualize if we first consider what the data look like prior to being summarized in a contingency table. For the Klondike volunteers, gender and race are coded as follows:

Sex	Race
1 = female	1 = white
2 = male	2 = nonwhite

Both variables are measured at the nominal level. Table 15.7 presents the raw data for 12 volunteers at the Art Museum.

The first row of data indicates that the first volunteer at the Art Museum has a score of "1" for both variables. This indicates that the volunteer is a white

Table 15.7 Data for 12 Volunteers at the Klondike Expressionist Art Museum

Sex	Race
1	1
2	2
1	2
2	2
2	1
1	1
1	2
1	1
2	2
1	2
2	2
2	1

female. The second row of data has values of "2" for both variables. This indicates that the second volunteer is a nonwhite male. To make sure you understand the pairings for sex and race, interpret the entries for volunteers 3 through 12.

The cross-tabulation of these two variables for all 451 volunteers at the Art Museum displays the number of cases (volunteers) that fall into each of the race–sex combinations. In this sample of museum volunteers composed of 142 white males, 67 white females, 109 nonwhite males, and 133 nonwhite females, we obtain the contingency table displayed in Table 15.8. This type of table is called a *cross-tabulation* because it crosses (and tabulates) each of the categories of one variable with each of the categories of a second variable.

The numbers in each cell category simply represent the aggregate results compiled from all 451 rows of data on volunteers, where each row corresponds to a volunteer. Although the cells within contingency tables such as Table 15.8 sometimes contain large numbers that may "look" like interval-level data, you should remember that these numbers represent total case counts for nominal- or ordinal-level variables. The data used to generate Table 15.8 look just like the data for the 12 volunteers displayed in Table 15.7, except that the data for all 451 rows (volunteers) are counted and summarized in Table 15.8.

At this point, some terminology is useful. The cross-classifications of the two variables—white-male, white-female, nonwhite-male, nonwhite-female—are called the **cells** of the table. The cell frequencies indicate the number of cases fitting the description specified by the categories of the row and column variables. The total number of respondents who are white or nonwhite is presented at the foot of the "White" and "Nonwhite" columns, respectively. Similarly, the total number of respondents who are male or female is presented at the far right of the respective rows. In reference to their position around the perimeter of the table, these total frequencies are called **marginals** (or marginal frequencies). These totals are calculated by adding the frequencies in the appropriate column or row. Finally, the **grand total**—the total number of cases represented in the table (**N**)—is displayed conventionally in the lower right corner of the table. It can be found by adding the cell frequencies, or the row marginals, or the column marginals. You should satisfy yourself that all three of these additions give the same result. You should also make certain that you understand what each number in Table 15.8 means.

Table 15.8 **Contingency Table: Race and Sex of Volunteers to Klondike Expressionist Art Museum**

	Race		
Sex	White	Nonwhite	Total
Male	142	109	51
Female	67	133	200
Total	209	242	451

Table 15.9 Relationship between Type of Employment and Attitude toward Balancing the Federal Budget

	Type of Employment			
Attitude toward Budget Balancing	Public	Private	Nonprofit	Total
Disapprove				
Approve				
Total				

Table 15.10 Relationship between Educational Level and Performance on Civil Service Examination

	Education		
Performance on Civil Service Examination	High School or Less	More Than High School	Total
Low	100	200	300
High	150	800	950
Total	250	1,000	1,250

To ensure that you can assemble a cross-tabulation, fill in the cell, marginal, and grand total frequencies in Table 15.9. The variables of interest are "type of employment" (public sector, private sector, or nonprofit sector) and "attitude toward balancing the federal budget" (disapprove or approve). The cell frequencies are as follows: public-disapprove 126; public-approve 54; private-disapprove 51; private-approve 97; nonprofit-disapprove 25; nonprofit-approve 38.

Relationships between Variables

Researchers assemble and examine cross-tabulations because they are interested in the relationship between two ordinal- or nominal-level variables. A **statistical relationship** may be defined as a recognizable pattern of change in one variable as the other variable changes. In particular, the type of question that is usually asked is, As one variable increases in value, does the other also increase? Or, as one variable increases, does the other decrease?

The cell frequencies of a cross-tabulation provide some information regarding whether changes in one variable are associated statistically with (related to) changes in the other variable. The cross-tabulation presented in Table 15.10 of "education" (high school or less; more than high school) with "performance on the civil service examination" (low; high) illustrates this idea.

At first glance, the table seems to indicate that as education *increases* from high school or less ("low") to more than high school ("high"), performance on

the civil service examination *decreases*, for twice as many individuals with high education (200) as individuals with low education (100) received low scores on the test. Because we would anticipate that education would *improve* scores on the examination, this initial finding seems counterintuitive. In fact, it is not only counterintuitive but also incorrect.

The reason for the faulty interpretation is that we have failed to take into account the *total number* of individuals who have low as compared with high education (that is, the marginal totals). Note that although this sample contained only 250 people with a high school education or less, 1,000 individuals—four times as many—had more than a high school education. Thus, when these figures are put in perspective, there are *four* times as many people with high education than with low education in the sample—yet only *twice* as many of the former as of the latter got low scores on the civil service examination. These data suggest that, in contrast to our initial interpretation of the table, more highly educated people do earn higher scores on the civil service examination than do the less educated. This finding accords with intuition and is the primary conclusion supported by the table—when one analyzes it correctly.

How does one do so? The analysis process has three major steps. The problem with the initial interpretation of the contingency table was that it overlooked the relative number of cases in the categories of education (that is, the marginal totals). This problem can be remedied by percentaging the table appropriately, which is the key to analyzing and understanding cross-tabulations. The steps in the analysis process are as follows:

Step 1: Determine which variable is *independent* and which is *dependent.* As explained in Chapter 3, the independent variable is the anticipated causal variable, the one that is supposed to lead to changes or effects in the dependent or response (criterion) variable. In the current example of the relationship between education and performance on the civil service examination, it is expected that higher education leads to improved performance on the test. Stated as a hypothesis: The higher the education, the higher the expected score on the civil service examination. Hence, education is the independent variable, and performance on the civil service examination is the dependent variable.

Step 2: Calculate percentages within the categories of the *independent* variable—in this case, education. We would like to know the percentage of people with high school education or less (low education) who received high scores on the civil service examination and the percentage of people with more than a high school education (high education) who received high scores. Then it would be possible to compare these percentages in order to determine whether those with high education receive higher scores on the examination than do those with low education. This comparison allows us to evaluate, on the basis of the data, whether the expectation or hypothesis stated previously is correct—that is, that education leads to improved scores on the civil service examination.

The procedure used to calculate percentages within the categories of education is the same as the univariate procedure elaborated earlier in the chapter. We are interested first in the percentage of people with high school education or less who received high scores on the civil service examination. Table 15.10 indicates that a total of 250 people fall into this category of education, and of these, 150 received high scores on the test. Thus, we find that $(150 \div 250) \times 100 = 60\%$ of those with low education earned high scores on the civil service examination. (Note that this is also the probability of receiving a high score on the exam given low education; see Chapter 7.) The other 100 of the 250 people with low education received low scores on the test; converting to a percentage, we find that $(100 \div 250) \times 100 = 40\%$ of those with low education earned low test scores (the probability of a low test score given low education).

Moving to those with more than a high school education, Table 15.10 shows that 800 of the 1,000 people with this level of education—or 80% [$(800 \div 1{,}000) \times 100$]—received high scores on the civil service examination, and the other 200—or 20% [$(200 \div 1{,}000) \times 100$]—earned low scores. Given high education, the probability of receiving a high score on the civil service examination is .80, and the probability of receiving a low score is .20 (see Chapter 7). All percentages have now been calculated. Table 15.11 presents the cross-tabulation percentaged within the categories of education, including all calculations.

Step 3: Compare the percentages calculated within the categories of the *independent variable* (education) for *one* of the categories of the *dependent variable* (performance on the civil service examination). For example, whereas 80% of those with high education earned high scores on the civil service examination, only 60% of those with low education did so. Thus, our hypothesis is supported by these data: In general, those with high education received higher scores on the examination than did those with low education. As hypothesized, the higher the education, the higher is the score on the civil service examination.

To summarize the relationship between two variables in a cross-tabulation, researchers often calculate a **percentage difference** across one of the categories

Table 15.11 Percentage Distribution for Data of Table 15.10

Performance on Civil Service Examination	Education	
	High School or Less	More Than High School
Low	$(100 \div 250) \times 100 = 40\%$	$(200 \div 1{,}000) \times 100 = 20\%$
High	$(150 \div 250) \times 100 = 60\%$	$(800 \div 1{,}000) \times 100 = 80\%$
Total	$(n = 250)$ 100%	$(n = 1{,}000)$ 100%

of the dependent variable. In this case, the percentage difference is equal to 80% minus 60%, or 20 percentage points (the percentage of those with high education who earned high scores on the test minus the percentage of those with low education who did so). The conclusion, then, is that education appears to make a difference of 20 percentage points in performance on the civil service examination. As you will learn in Chapter 16, the percentage difference is a measure of the strength of the relationship between two variables.

Example: Automobile Maintenance in Berrysville

The city council of Berrysville, California, has been under considerable pressure to economize. Last year, the council passed an ordinance authorizing an experimental program for the maintenance of city-owned vehicles. The bill stipulates that for 1 year, a random sample of 150 of the city's 400 automobiles will receive no preventive maintenance and will simply be driven until they break down. The other 250 automobiles will receive regularly scheduled preventive maintenance. The council is interested in whether the expensive program of preventive maintenance actually reduces the number of breakdowns. After a year under the experimental maintenance program, the city council was presented with the data in Table 15.12, which summarizes the number of automobile breakdowns under the no maintenance and preventive maintenance conditions. Analyze the data for the city council, and help the council by making a recommendation regarding whether the program should be continued (and/or expanded) or terminated.

Step 1: Determine which variable is independent and which is dependent. There should be no doubt that automobile maintenance is expected to affect the number of breakdowns. Therefore, "maintenance" is the independent variable, and "breakdowns" is the dependent variable. Stated as a hypothesis, we have the following: The greater the level of maintenance, the less the rate of breakdowns.

Step 2: Calculate percentages within the categories of the independent variable, "automobile maintenance." The calculations are shown in Table 15.13.

Step 3: Compare percentages for one of the categories of the dependent variable. More than half (52%) of the automobiles that received no maintenance

Table 15.12 Automobile Maintenance Data

	Automobile Maintenance		
Automobile Breakdowns	None	Regularly Scheduled	Total
No breakdown	72	194	266
Breakdown	78	56	134
Total	150	250	400

Table 15.13 Percentage Distribution for Data of Table 15.12

Automobile Breakdowns	Automobile Maintenance: None	Automobile Maintenance: Regularly Scheduled
No breakdown	$(72 \div 150) \times 100 = 48\%$	$(194 \div 250) \times 100 = 77.6\%$
Breakdown	$(78 \div 150) \times 100 = 52\%$	$(56 \div 250) \times 100 = 22.4\%$
Total	$(n = 150)$ 100%	$(n = 250)$ 100.0%

broke down during the 1-year experimental program, compared to just 22.4% of the automobiles that received regularly scheduled maintenance. This is a difference of 29.6% (52% – 22.4%). Thus, automobile maintenance appears to make nearly a 30% difference in the rate of breakdowns. The data show support for the hypothesis: As maintenance increases, the number of breakdowns decreases by almost 30%. From these data, should you recommend to the city council that it continues or terminates the experimental maintenance program?

Note: When these data were released to the public, the Berrysville press made great sport of the folly of the city council for experimenting with the "dang fool" (no) maintenance program. The members of the city council who had voted for the program were soundly defeated in the next election. In the first meeting of the new city council, the researcher who had compiled and analyzed the automobile maintenance data was awarded a substantial raise in salary. There may be a moral to this story.

Larger Contingency Tables

With a single exception, the examples of contingency tables presented in this chapter have consisted of "two-by-two" tables—cross-tabulations in which both the independent and the dependent variables comprise just two values or response categories. Cross-tabulations can and often do consist of variables with a greater number of response categories. For example, Table 15.14 presents the cross-tabulation of "income" (low, medium, and high) and "job satisfaction" (low, medium, and high)—How satisfied are you with your job?—for the employees of the Maslow City Post Office.

Although the analysis becomes more complicated, contingency tables based on variables with many response categories are analyzed in the same way as are the smaller two-by-two tables. Start by determining which variable is independent and which is dependent. In this example, you would expect income to lead to job satisfaction: The higher the income, the higher would be the expected job satisfaction. "Income" is the independent variable, and "job satisfaction" is the dependent variable. Therefore, the table should be percentaged within the categories of

Table 15.14 Relationship between Income and Job Satisfaction

	Income			
Job Satisfaction	Low	Medium	High	Total
Low	100	30	10	140
Medium	60	80	15	155
High	40	40	50	130
Total	200	150	75	425

income. Table 15.15 presents the cross-tabulation, percentaged according to the steps elaborated earlier.

The final step in the analysis of contingency tables is to compare percentages for one of the categories of the dependent variable. Although the choice of a category in two-by-two tables is not a critical decision—both categories of the dependent variable will yield the *same* percentage difference—in larger tables, the selection of a category of the dependent variable for purposes of percentage comparison requires more care. In general, you should *not* choose an *intermediate* category, such as "medium" job satisfaction, for this purpose. Choosing either of the *endpoint* categories—"low" or "high" job satisfaction—will result in clearer understanding and interpretation of the contingency table.

Once the (endpoint) category of the dependent variable has been selected, compare the percentages calculated for the *endpoint* categories of the independent variable. Again, avoid intermediate categories of the independent (and dependent) variable for this purpose. In Table 15.15, this rule suggests that we compare the percentage of those with low income who have high job satisfaction (20%) with the percentage of those with high income who have high job satisfaction (66.7%). Alternatively, we could compare the percentage of those with low income who express low job satisfaction (50%) with the percentage of those with high income who express low job satisfaction (13.3%).

Table 15.15 Percentage Distribution for Data of Table 15.14

	Income		
Job Satisfaction	Low	Medium	High
Low	$(100 \div 200) \times 100 = 50\%$	$(30 \div 150) \times 100 = 20.0\%$	$(10 \div 75) \times 100 = 13.3\%$
Medium	$(60 \div 200) \times 100 = 30\%$	$(80 \div 150) \times 100 = 53.3\%$	$(15 \div 75) \times 100 = 20.0\%$
High	$(40 \div 200) \times 100 = 20\%$	$(40 \div 150) \times 100 = 26.7\%$	$(50 \div 75) \times 100 = 66.7\%$
Total	$(n = 200)$ 100%	$(n = 150)$ 100.0%	$(n = 75)$ 100.0%

Which percentage comparison(s) should the researcher use to summarize the relationship found in the cross-tabulation? The percentage difference calculation typically yields different results depending on the endpoint category of the dependent variable chosen. In the current case, the percentage difference based on high job satisfaction is 66.7% – 20.0% = 46.7%, whereas the percentage difference for low job satisfaction is 50.0% – 13.3% = 36.7%. These percentage differences suggest varying levels of support for the relationship.

Probably the best course of action for the public or nonprofit manager is to consider and report *both* figures. They show that those with high income indicated high job satisfaction more often than did those with low income (by 47%) and, conversely, that those with low income indicated low job satisfaction more often than did their counterparts with high income (by 37%). Thus, income appears to make a difference of 37 to 47 percentage points in job satisfaction. These figures provide support for the hypothesis that the greater the income, the greater is the expected job satisfaction. Chapter 16 presents other techniques especially appropriate for the analysis of larger contingency tables.

Displaying Contingency Tables

A set of conventions has been developed for presenting contingency tables. First, contingency tables are rarely presented as bivariate frequency distributions. Instead, you should display the table in percentaged form; the percentages should be calculated and displayed according to the procedures described in the preceding section (do *not* show the percentage calculations). Second, the independent variable is placed along the *columns* of the table, and the dependent variable is positioned down the *rows.* Third, the substantive meaning of the categories of the independent variable should show a progression from least to most moving from left to right across the columns, and the categories of the dependent variable should show the same type of progression moving down the rows. In other words, the categories should be listed in the order "low," "medium," and "high"; or "disapprove," "neutral," and "approve"; or "disagree," "neutral," and "agree"; and so on. This procedure greatly facilitates the interpretation of measures of association (discussed in Chapter 16). See Table 15.15 for an illustration. Fourth, the percentages calculated within categories of the independent variable are summed down the column, and the total for each category is placed at the foot of the respective column. The sum should equal 100%, but because of rounding error, it may vary between 99% and 101%. *Do not add the percentages across the rows of the table; this is a meaningless operation.* Finally, the total number of cases within each category of the independent variable is presented at the foot of the respective column. Usually, these totals are enclosed in parentheses and contain the notation $n =$ _____. Table 15.16 presents schematically a contingency table displayed according to the conventional rules.

Two problems arise regarding the conventional display of contingency tables. First, these rules are widely accepted—but not always. Thus, in reading and

Table 15.16 Conventional Format for a Contingency Table

Dependent Variable	Independent Variable			
Substantive meaning of categories increases	Substantive meaning of categories increases (e.g., "low," "medium," "high") →			
↓	——%	——%	——%	——%
	——%	——%	——%	——%
	⋮	⋮	⋮	⋮
Total	100.0%	100.0%	100.0%	100.0%
	(n = _____)	(n = _____)	(n = _____)	(n = _____)

studying contingency tables presented in books, journals, reports, memoranda, magazines, newspapers, and so on, you should not assume that the independent variable is always along the columns or that the dependent variable is always down the rows. Nor can you assume that the categories of the variables are ordered in the table according to the conventions. Instead, you should examine the table critically, decide which variable is independent and which is dependent, check to see whether the percentages have been calculated within the categories of the independent variable, and verify whether the author has compared percentages appropriately. You should recognize these procedures as the steps specified for analyzing and interpreting cross-tabulations presented in this chapter. Cultivating this habit will not only increase your understanding of contingency table results but also sharpen your analytical skills.

The second problem arises as a consequence of computer utilization. On the job (or in class), you may be dealing with contingency tables constructed and percentaged by a computer. Not only is the computer oblivious to the distinction between independent and dependent variables, the ordering of response categories of variables, and so on, but also computers are usually programmed to print out *three different sets of percentages*: percentages calculated (1) within the categories of the row variable; (2) within the categories of the column variable; and (3) according to the total number of cases represented in the contingency table, sometimes called *corner* or *total* percentaging. It is up to you as the data analyst to determine which set of percentages is most meaningful and, if necessary, to reconstruct the contingency table by hand from the computer printout according to the conventional form described here. If you follow the steps for the analysis of contingency tables developed in this chapter, this task should not be difficult.

This chapter has elaborated a general method for determining whether two variables measured at the nominal or ordinal level are related statistically: contingency table analysis or cross-tabulation. However, it has not addressed the

question of *how strongly* two variables are related. This question serves as the focus for the next chapter.

Chapter Summary

Contingency tables are used to display and analyze the relationship between two variables measured at the nominal or ordinal level. The simplest and often most useful technique for analyzing contingency tables is to calculate percentages appropriately and to compare them.

This chapter illustrated the analysis of contingency tables. A contingency table is a bivariate—or two-variable—frequency distribution. It presents the number of cases that fall into each possible pairing of the values of two variables. There are three major steps in the analysis process. First, determine which variable is independent and which is dependent. Second, calculate percentages within the categories of the independent variable. Finally, compare the percentages calculated within the categories of the independent variable for one of the categories of the dependent variable, and interpret the results. Contingency tables for variables with more than two response categories are analyzed using the same basic approach. You should calculate and report the percentage differences for both endpoint categories of the dependent variable.

The chapter concluded by presenting a standard format to display and analyze contingency tables. Although you cannot assume that all contingency tables are set up in this manner, closely examining all cross-tabulations will enhance and sharpen your analytical skills.

Problems

15.1 The Lebanon postmaster suspects that working on ziptronic machines is the cause of high absenteeism. More than 10 absences from work without business-related reasons is considered excessive absenteeism. A check of employee records shows that 26 of the 44 ziptronic operators had 10 or more absences and 35 of 120 nonziptronic workers had 10 or more absences. Construct a contingency table for the postmaster. Does the table support the postmaster's suspicion that working on ziptronic machines is related to high absenteeism?

15.2 During last year's budget crunch, several deserving employees of the Bureau of Procedures (BP) were denied promotions. This year, an unusual number of BP employees retired. The bureau chief suspects that the denial of promotions resulted in increased retirements. Of the 115 employees denied promotion, 32 retired. Of the 58 employees promoted, 9 retired. Present a contingency table, and analyze this information.

15.3 The Egyptian Air Force brass believe that overweight pilots have slow reaction times. They attribute the poor performance of their air force in recent war games

in the Sinai to overweight pilots. The accompanying data were collected for all pilots. Analyze these data for the Egyptian Air Force brass.

	Pilot Weight		
Reaction Time	Normal	Up to 10 Pounds Overweight	More Than 10 Pounds Overweight
Poor	14	36	45
Adequate	35	40	33
Excellent	46	25	15

15.4 Auditors for the Military Airlift Command (MAC) are checking the arrival times of the three charter airlines they used in the Pacific last year. Branflake Airways flew 135 flights and was late 78 times. Flying Armadillo Airlines flew 94 flights and was late 35 times. Air Idaho flew 115 flights, with 51 late arrivals. Set up a contingency table, and analyze it for MAC.

15.5 The state personnel office oversees the state's tuition assistance program, which pays the tuition of civil servants taking courses for an MPA. Only two schools offer an MPA degree in the state capital: Capital College of Law and East Winslow State University. Some concern is expressed by legislators that many tuition-assisted students do not graduate. Analyze the data in the accompanying table for the personnel office.

	Students Assisted for MPA Tuition	
Status	Capital	East Winslow
Did not graduate	69	83
Graduated	23	37

15.6 Hyram Drant, a research analyst for the city fire department, suspects that old water pumps are more likely to fail. From the data in the accompanying table, construct a contingency table and check Drant's suspicion. How else could this problem be analyzed?

Age of Pump in Years		
Pump Failed	Pump Did Not Fail	
23	15	7
47	6	9
11	9	4
53	33	19
26	26	36
15	17	47
42	9	31
37	12	23
	31	6
	46	9
	15	3

15.7 As head scheduler of special events for the Incomparable Myriad (the city arena), your task is to schedule events that make a profit so that the city need not subsidize the arena. Analyze the data in the accompanying table, which is based on last year's data, and write a report to the city council.

	Type of Event				
Status	Hockey Games	Religious Rallies	Basketball Games	Rock Concerts	Public Administration Conventions
Not profitable	24	4	21	2	3
Profitable	18	32	6	8	0

15.8 As the newly appointed head of evaluation for the state agriculture experiment station, you are asked to evaluate the relative effectiveness of corn hybrids AX147 and AQ49. Of 32 test plots, AX147 had high yields on 21. AQ49 had high yields on 17 of 28 test plots. Construct a contingency table and make a recommendation.

15.9 The Cancer Institute is evaluating an experimental drug for controlling lip cancer. Eighty lip cancer victims are randomly selected and given the drug for 1 year. Sixty other lip cancer victims are randomly selected and given a placebo for a year. From the data in the accompanying table, what would you conclude?

	Treatment Group	
Cancer Status	Drug Group	Placebo Group
Active	58	42
Remission	22	18

15.10 A supervisor in the Department of Rehabilitative Services is critical of the performance of one of her counselors. The counselor is expected to arrange job training for those in need of vocational rehabilitation so that they may find employment. Yet the counselor has managed to place just 35% of his clients. The counselor argues that he is actually doing a good job and that the reason for his overall low rate of placement is that most of his clients are severely disabled, which makes them very difficult to place. The counselor's case load is presented in the accompanying table. Percentage the table appropriately, and evaluate who is correct—the supervisor or the counselor.

	Clients	
Job Placement	Not Severely Disabled	Severely Disabled
Not placed	17	118
Placed	47	26

15.11 A professor of public administration has kept records on the class participation of his students over the past several years. He has a strong feeling (hypothesis) that class participation is related to grade in the course. For this analysis he classifies course grades into two categories, fail and pass. He operationalizes class

participation as "low" if the student participated in class discussion in fewer than 25% of class periods, and "high" if the student participated in 25% or more of the periods. Based on these definitions, he has assembled the cross-tabulation below. Does a relationship exist between class participation and course grades?

	Class Participation	
Grade in Course	Low	High
Fail	56	15
Pass	178	107

15.12 Susan Wolch and John Komer are interested in determining which of two books is more effective in teaching statistics to public administration students. They randomly assign a pool of 50 students to two groups of 25 students each. One group uses Meier, Brudney, and Bohte, *Applied Statistics for Public and Nonprofit Administration.* The other group uses Brand *X.* Their criterion for measuring success is student grades in the course. They get the results shown in the accompanying table. Evaluate these data and make a recommendation.

	Book Used in Class	
Grade	Brand X	Meier, Brudney, and Bohte
Students receiving C's, D's, or F's	18	9
Students receiving A's or B's	7	16

15.13 Madonna Lewis's job in the Department of Sanitary Engineering is to determine whether new refuse collection procedures have improved the public's perception of the department. A public opinion survey was taken both before and after the new procedures were implemented. The results appear in the accompanying table. Analyze the table, and evaluate whether public perception of the department has improved over time.

	Survey	
Opinion	Before	After
Department is doing a poor job	79	73
Department is doing a good job	23	47