ANALYZING DATA

Bivariate Relationships

7

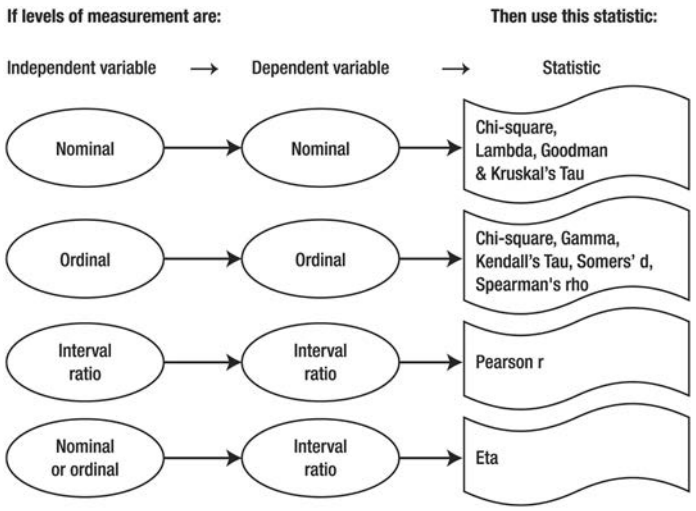
The invalid assumption that correlation implies cause is probably among the two or three most serious and common errors of human reasoning.

—Stephen Jay Gould, scientist

LEARNING GOALS

Understanding bivariate statistical analysis is the focus of this chapter. Central to this is learning how to read and construct cross-tables of data and deciding which statistics to use to measure association and correlation (see Figure 7.1). By the end of the chapter, you should understand how to reject or accept a hypothesis using the appropriate statistics to assess bivariate relationships. You should also be able to put together cross-tables and interpret them clearly in words.

Figure 7.1 Statistical Decision Steps (also see Statistical Analysis Decision Tree in Appendix)



After you evaluate your univariate data analysis and feel confident that you have variables with good distributions, it is now time to begin investigating the bivariate relationships you proposed in your research questions and hypotheses. *Bivariate data analysis* assesses the association between two variables. If you are interested in studying three or more variables at a time, then multivariate analysis is required (see Chapter 9). Remember, if for some reason any of your variables approximate a constant—that is, almost everyone has selected only one or two of the values for that variable—then you must eliminate them from further analysis. For example, when 90 percent of the people who completed the questionnaire agree or strongly agree that they are satisfied with their current job, then job satisfaction may no longer be a variable in your study unless the sample size is very large.

Keep in mind what the objective is: to demonstrate whether the independent variable is exerting significant influence on the dependent variable in order to infer something about a population from which your sample was selected. You want to see if the variation that exists among the respondents in the dependent or outcome variable can be explained or predicted by the variation that exists in the independent variable. For example, you might want to establish whether there is a relationship between spending time posting photos on Instagram and earning lower or higher grades in school. When there is some statistically significant relationship, you have fulfilled one of the three requirements to establish cause and effect (see Chapter 1). Once you have eliminated alternative explanations and can show that the independent variable occurred prior in time to the dependent variable, then you tweet to everyone that you have found a cause for the occurrence of the dependent variable.

Another objective may simply be to describe how the dependent variable varies according to different categories of the independent variable. You might be interested in showing whether political party affiliation differs between men and women or among Hispanic, Asian/Pacific Island people, and African Americans. Or you would like to report how many respondents in your study are young, middle-age, and elderly and who are from rural and urban areas, such as young urban, young rural, elderly urban, elderly rural, and so on. This chapter focuses on developing skills for accomplishing these objectives through bivariate data analysis and for understanding measures of correlation and association.

PRESENTING NOMINAL AND ORDINAL DATA IN TABLES

Using your research questions and hypotheses, begin by making a list of the variables in them. Next, label each variable in terms of its level of measurement (nominal, ordinal, or interval/ratio), and—on the basis of what follows in this and the next two

chapters—decide the appropriate way of presenting the data visually with tables or graphs. Finally, select the most relevant statistics to assess what you have found.

Similar to the way you show the distribution of a single variable, constructing frequency tables for both variables simultaneously is a good way to start analyzing the variables in your hypotheses. What is the occurrence of one variable in terms of the other one? Is one variable's distribution contingent on the other's variation? Is the dependent or outcome variable associated with the independent or predictor variable in any way? To answer these kinds of questions visually, a table crossing the two variables is made. These are usually referred to as *contingency tables* or *cross-tabulations* (*crosstabs*, for short). They illustrate the number and percentage of occurrences in the sample of each value of one variable simultaneously with each value of the other variable. Remember, variables are usually the questions in your survey, like "age"; values are the answers to the questions, like "under 25" or "between 40 and 50."

Crosstabs are ideally suited to nominal or ordinal measured variables or to interval/ratio data with a very limited number of discrete values. Each value (or category) of one variable constitutes the columns in the table, and each value (or category) of the other variable makes up the rows of the table. It is easy to remember which is which: Just as *columns* hold up a building, they hold up a table; just as you move across the *row* to get to your seat in the movie theater, you move across the rows in a table. These are similar to the concepts used in a spreadsheet program, like Excel, or a data analysis software program, like SPSS.

Ideally, the number of rows and columns in a table should be small enough to see all of them at the same time on a screen or sheet of paper, thereby restricting crosstabs to mostly nominal or ordinal measures. Interval/ratio data with a limited number of values can also be used, but imagine a table that has a row for every single age that exists in your sample of 18- to 89-year-olds! The size of a table is referred to in terms of its rows and columns, in that order, such as a 3 × 2 table (three rows and two columns). *Cells* are the locations where each row's values and each column's values intersect, and the number of them is quickly calculated by multiplying the number of rows by the number of columns: A 3 × 2 table has six cells, for example. A cell's location is also given in terms of its row and column number, so you can refer to the data in a cell as being "in row 3, column 1."

Another convention is that the percentages for each value (or category) of the independent variable should add up to 100 percent; if the values of the independent variable are columns, as is recommended, then each column should total 100 percent. In this way, you standardize the responses when the numbers of responses for each value of the variable are not equal. This allows you to make a fair comparison about the frequency of occurrence of each value of the independent variable in terms of each value of the dependent variable.

Let's say you are interested in seeing whether there is some relationship between gender and political party affiliation. Your hypothesis is two-tailed, so you simply state that there is a relationship between the two, or using the null form, there is no relationship between gender and political party affiliation. It seems logical to see whether political party choice is an outcome of gender (clearly, choosing a party doesn't result in your gender); in other words, political party affiliation depends on the independent variable of gender. The values for gender (male and female) compose the columns, and the values for party (Republican, Democrat, Independent, Other) become the rows.

Table 7.1, a bivariate frequency table or cross-tabulation (*crosstab*), shows that there are 68 women who are Democrats in this sample. Remember, data are based on valid responses, and to be included, respondents had to answer both questions, with no blanks or missing values. Because percentage is used to standardize unequal category numbers, we can compare 69.4 percent with 37.7 percent. Notice that although there are 68 women and 20 men who are Democrats, it would not be correct to say that more than three times as many women are Democrats than men. This would assume that the same number of men and women were answering the questions,

Table 7.1 Bivariate Cross-Table Example, SPSS

			GENDER		
			Female	Male	Total
Political Party	Democrats	Count	68	20	88
		% of political party	77.3%	22.7%	100.0%
		% of gender	69.4%	37.7%	58.3%
	Republicans	Count	8	15	23
		% of political party	34.8%	65.2%	100.0%
		% of gender	8.2%	28.3%	15.2%
	Independents	Count	18	13	31
		% of political party	58.1%	41.9%	100.0%
		% of gender	18.4%	24.5%	20.5%
	Other	Count	4	5	9
		% of political party	44.4%	55.6%	100.0%
		% of gender	4.1%	9.4%	6.0%
Total		Count	98	53	151
		% of political party	64.9%	35.1%	100.0%
		% of gender	100.0%	100.0%	100.0%

but as you can see, the total sample is 98 women and 53 men. To make up for the unequal totals for each value of the independent variable, a percentage is used instead to make the comparison. It would be more accurate to say that approximately twice as many women as men are Democrats, in terms of the percentage that is relative to their sample sizes.

The total numbers at the bottom of the columns and at the right end of the rows are called *marginals*, because they appear at the edges or margins of the table. They essentially provide you the same information you would have gotten if you had constructed two separate univariate frequency tables. Note that this table includes percentages in terms of the dependent, or row, variable in addition to column percentages. For example, the 13 respondents who are male and registered as Independents can be read either as 24.5 percent of the men are Independents, or you can say that 41.9 percent of all Independents in your study are male.

These are two different things; it all depends on how you want to report your data. If someone from the Republican Party wanted to know the breakdown of men and women in the party, then you would use the row percentages (percent of political party). If, on the other hand, you wanted to know whether there was a difference between men and women in terms of Republican affiliation, then use the column percentages (percent of gender). Sometimes there is no clear-cut independent (cause) and dependent (effect) variable because both can occur at the same time. These are often referred to as *symmetrical* relationships because either can be seen as the predictor for the other and either one can be the outcome needing explanation.

This can be confusing, so think aloud about what you really want to know and figure out how to show the percentages according to your objectives. It is not the same thing to say that 20 percent of women on campus major in political science and that 20 percent of political science majors are women. These are two different meanings, and it is crucial to make the distinction when presenting your findings in bivariate cross-tabulations.

A good rule of thumb is the values of a variable that you want to compare should add up to 100 percent. If you want to compare men with women and which majors they choose, the two categories (values) of gender should each add up to 100 percent. Then you can say what percentage of men select political science compared to what percentage of women do, and so on. But if you want to compare the majors and find out how many men and women are in each major, the categories for major (political science, sociology, psychology, etc.) should each add up to 100 percent. Then you can find out what percentage of political science majors are female and male compared with the percentage of sociology majors who are female and male, and so on.

TESTING BIVARIATE RELATIONSHIPS

Crosstabs illustrate whether some relationship is occurring with your data. There seems in Table 7.1 to be some difference in political party affiliation between men and women. Sometimes, however, the percentages and frequencies are so close that it is not possible to tell if there is a meaningful difference at a glance. Even when there appears to be a difference, you still need a more objective assessment than your own, perhaps selective, perspective to tell you whether the relationship is a significant one and not something that could occur just by chance.

Chi-Square

Rest assured, there is a way of testing whether there is a significant relationship between your variables! One of the most important and frequently used statistics to assess the association between ordinal and nominal measures is called *chi-square* (χ^2, pronounced "kie-square"). This statistic measures how independent your two variables are and asks whether what you found (observed) is significantly different from what you would have expected to get by chance alone. The calculation looks at each cell and measures the difference between the actual frequency you got and the frequency that would have been expected by chance. For example, if half the sample is male and half are female, and half are Hispanic and half are Asian, then you would expect by chance 25 percent of the sample to be in each cell in that 2 × 2 table; that is, 25 percent should be Hispanic men, 25 percent should be Asian women, and so on.

Like the concept of the standard deviation, a mean deviation for each cell is calculated, and these are added together to produce a number called chi-square. Obviously, when there are many values for both variables, the larger the final number will be since there are many more cells to add together. This number is compared by the data analysis software program to a sampling distribution of chi-square values; this chi-square distribution approaches a normal curve when the number of cells increases, or more accurately when the degrees of freedom increase (see Box 7.1). The chi-square number in itself is not a measure of strength or magnitude; it must be compared to a distribution of chi-square values and not to other chi-squares unless the sample sizes and number of cells are the same. Its value depends on the size of the sample and the number of cells: The more cells there are, the larger the chi-square value is likely to be.

You cannot tell whether it is statistically significant by simply looking at the chi-square number. The probability of obtaining a chi-square value by chance for a particular number of cells in a table is determined by the computer program that compares the value to a normal curve table of probabilities. If the probability is less than .05 (or whatever alpha level you set), then you can declare there is an association between your two variables by rejecting the null hypothesis of no association.

For the crosstab in Table 7.1 on political party and gender, a chi-square value of 17.361 was calculated for these data and is shown in Table 7.2. Notice there are three degrees of freedom (*df*) since there are four political party rows (4 – 1 = 3) and two gender columns (2 – 1 = 1), resulting in 3 × 1 = 3. What this statistical test tells us is that for a table with three degrees of freedom and 151 respondents, a chi-square value of 17.361 is significant at the .001 level for a two-tailed hypothesis. The probability of obtaining a chi-square value of 17.361 by chance alone is less than 1 in 1,000 ($p < .001$). We therefore reject the null (that there is no difference in political party affiliation between men and women) and accept the alternative hypothesis that there is an association between gender and political party preference. We conclude that women in this particular sample are statistically more likely to be registered Democrats than are men. It's not enough just to say there is an association—you must also say what it is, or else your Facebook sharing will be fairly dull! Remember also that you are talking about a collection of men and women, not about any one particular person. Although you can say that women are more likely to be registered Democrats in this sample, you cannot generalize to all women (unless this is a random sample, which it is not), and you cannot say that any individual woman will be a registered Democrat.

The value of the chi-square statistic is also affected by cells with low frequencies, hence the information provided in footnote "a" that lets you know how many cells have small numbers. A rule of thumb is that every cell should have at least five expected respondents; you can see now how the size of the sample can make a difference when you have a table that is large, say five racial/ethnic groups and five religions, resulting in a 25-cell table. If your table has small numbers in its cells, there are statistical corrections, such as Fisher's Exact Test. If you have a 2 × 2 table, Yates's Correction should be used to adjust the calculation of chi-square. More information about these statistics can be found in advanced statistics books, on the Internet, and in most computer statistical programs.

Table 7.2 Example of a Chi-Square Test, SPSS

CHI-SQUARE TEST			
	Value	*df*	Asymptotic Significance (two-tailed)
Pearson chi-square	17.361[a]	3	.001
N of valid cases	151		

[a] One cell (12.5 percent) has expected count less than 5. The minimum expected count is 3.16.

BOX 7.1
CALCULATING CHI-SQUARE AND DEGREES OF FREEDOM

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Calculating a chi-square by hand is fairly simple. For each cell in the table, take what is actually found in the study or the observed value (*O*) and subtract from it the expected value (*E*). Figuring out what is expected in each cell is the tricky part. This is calculated in several steps: (1) Take the total number of people in a column and divide that by the total of all respondents answering the questions in order to get what percentage of the entire valid sample holds the value of the variable in the first column; (2) then use that percentage and multiply it by the total row value for each row associated with the first column. Next, go on to the second column, take its total and divide by the grand total, and use that percentage and multiply it by the row total for each row in the second column, and so on. Using a formula, it's

$$\frac{\text{Column total}}{\text{Total total}} \times \text{Row total}$$

Imagine these are your results in raw numbers:

	MEN	WOMEN	TOTALS
Psychology	28	56	84
Sociology	26	35	61
Political science	67	27	94
Totals	121	118	239

1. Take the column total for men and divide by the total of all those answering the questions: 121/239 = 0.506 or 50.6%
2. Multiply that percentage by the row total for each row in the men column to get the expected:

   50.6% × 84 = 42.5
   50.6% × 61 = 30.9
   50.6% × 94 = 47.6

Note these numbers add up to 121, the column total. All we did was say that if men make up 50.6 percent of the total number of people answering these two questions, then we would expect 50.6 percent of psychology majors to be men, and 50.6 percent of sociology majors, and 50.6 percent of political science majors. Since there are 84 psych majors, we would expect 50.6 percent of 84 to be men, that is, 42.5, and so on for each major.

BOX 7.1 CONTINUED

In each cell, the expected frequency is then subtracted from the actual observed one, that number is squared, and the result is divided by the expected frequency. Do this for each cell, and sum all those numbers to arrive at chi-square.

In our example, the observed for the first cell is 28 and the expected is 42.5. Already we can see that there is a deviation from the expected. There is an underrepresentation of men who are psych majors.

1. First, we subtract the expected from the observed:
   $28 - 42.5 = -14.5$
2. Then, we square the difference; recall how we did something similar to get the standard deviation—we squared the deviations of each score from the mean, which is the expected:
   $-14.5^2 = (-14.5) \times (-14.5) = 210.25$
3. Next, we divide this number by the expected:
   $210.25/42.5 = 4.95$
4. We continue to do this for each cell:
   $26 - 30.9 = -4.9$
   $4.9^2 = 24.01$
   $24.01/30.9 = 0.78$ and so on for each cell.
5. We add up all the calculations for each cell to arrive at chi-square:
   $4.95 + 0.78 + 7.91 + 5.07 + 0.80 + 8.11 = 27.62$

The chi-square value is compared to a computer-generated distribution of chi-squares to find the probability of achieving that chi-square by chance. Should it be less than .05 (or whatever standard was set such as .01 or .001), then you declare a statistically significant finding and hold your press conference or post your tweet!

Remember the size itself tells you nothing unless you know the number of cells. In this example, 27.62 is not necessarily statistically significant just because it is a particular value. A chi-square with a value of 8.5, for example, might be significant and one of 30 might not. It all depends on the number of cells and the size of the sample.

Degrees of Freedom

In this example, we have a 3 × 2 table, or six cells. Probability levels make use of a concept called degrees of freedom. *Degrees of freedom* (*df*) are the number of values that can vary in any calculation. When there is a fixed outcome (such as the sum of deviations around a mean always equals zero), it limits how many numbers can contribute to that outcome, usually by one ($N - 1$). For example, consider the following: The total number of seats in a theater row was ten, and three groups wanted to be seated. A party of four enters, then a group of two; only four more people can now be accommodated. The first two numbers, four and two, could have been any size—they were free to vary depending on which group got there first. Once established, however, the last number is fixed. If a party of five came first, then three more came along, there is now only room for a group of two. Out of three groups of people, two are free to vary, and only one group is fixed, or $N - 1$, that is, $3 - 1 = 2$.

BOX 7.1 CONTINUED

For cells in a crosstab, the degrees of freedom are the number of rows minus one times the number of columns minus one or $(r - 1) \times (c - 1)$. In this example, it would be two rows (3 – 1) and one column (2 – 1), for a total of two degrees of freedom.

In fact, if you calculated the expected frequencies for the first cell (row 1, column 1) and then for the second cell in the example (row 2, column 1), you didn't have to calculate any more. All you had to do was subtract the expected in the first cell (42.5) from the row total of 84 to figure out the expected for the cell in row 1, column 2. It has to add up to 84, so the expected for that cell must be 41.5. That cell's value is not free to vary once you calculated the first cell's. Notice that with two numbers adding up to 84 one is fixed and one is free to vary ($df = N - 1$).

Similarly, add up the expected frequency from cell 1 with the one below it (30.9) and subtract that number from the column total of 121 to get the expected value of the cell at row 3, column 1: 42.5 + 30.9 = 73.4, then 121 – 73.4 = 47.6. In short, only two cells are free to vary and had to be calculated; the rest are not free to vary. Again, the degrees of freedom are $N - 1$; in this case, three rows minus one results in two. Out of six cells, there are two degrees of freedom and four are fixed. Degrees of freedom are also used as a correction when estimating a population parameter from a sample statistic, as was seen in Chapter 6 when discussing the formula to calculate the standard deviation with its denominator of $N - 1$.

Correlation Coefficients

Chi-square is not the best statistic to use if you want to compare the results with other data because it is not a measure of comparative strength. If you are interested in the *strength* of an association, then correlation coefficients must be calculated. There are many available depending on the level of measurement for the variables being assessed. *Coefficients* range in value from 0.0 to 1.0. Like all coefficients, the closer a coefficient is to 1.0, the stronger the association. The closer a coefficient is to 0.0, the weaker the relationship. Except for nominal data where there is no order, coefficients also have either a negative or positive direction. A relationship is an *inverse* or *negative* one when an increase in one variable goes along with a decrease in the other; it's a *positive* relationship when a decrease in one variable goes along with a decrease in the other, or an increase in one goes with an increase in the other.

The strength of the correlation is not weakened in any way with the appearance of a minus sign. A correlation of −0.54 is stronger than a positive correlation of 0.45, for example. Sometimes the direction is due simply to the way the variables were measured; that is, if one variable is measured with 1 = strongly agree and 5 = strongly disagree, and another variable has 1 = strongly disagree and 5 = strongly agree, then a negative correlation between two variables measuring similar traits could result.

An easy way to understand coefficients is to think about them as the amount of change you have in your pocket. The closer the number is to $1.00, the more you have. If your coefficient is 0.83, then you have 83 cents, almost a dollar. If you get a coefficient

of 0.17, then you have small change. Roughly, coefficients below 0.30 may be considered weak, those between 0.30 and 0.70 may be moderate, and those above 0.70 are usually strong. But it's all relative in comparison to other studies and the size of the sample: If most research has found a coefficient of 0.14 between religious affiliation and political party choice, and you get 0.35, then yours might be considered fairly strong.

When evaluating a coefficient and its significance level, be sure to make a distinction about which numbers you are reading. Correlation coefficients range from 0.0 to 1.0, with the larger number meaning a strong and, very likely, statistically significant relationship. On the other hand, probability levels also range from 0.0 to 1.0, but smaller numbers mean low probability of chance occurrences. Because the goal is to get a statistic that did not happen by chance a good deal of the time—it is a finding due to the actual influence of the independent variable on the dependent variable, not due to chance—then the objective is to have a probability level that is low: $p < .05$, $p < .01$, or $p < .001$, the typical numbers used as cutoff points for significance levels. The closer the probability level is to 0 (that is, no probability the statistic occurred by chance), the more likely that what you found did not occur accidentally. The closer the correlation coefficient is to 1.0, the stronger it is; the independent variable explains almost all the variation in the dependent variable. You will make fewer errors in explaining or predicting it. So be alert to which number is the correlation coefficient and which one is the probability level.

It is important to know, though, that low correlation coefficients can be statistically significant with large sample sizes; thus, it is better to look at the strength of correlations rather than to focus only on their significance levels.

Measures of Association: Nominal Variables

Although chi-square is the statistic most used to test the association between two variables, many other statistics are available to assess the strength of a relationship. When you are interested in evaluating relationships between two *nominal* level variables, then use such statistics as the *Phi* (ϕ) *coefficient* (for 2 × 2 tables), the *Contingency coefficient* (for larger than 2 × 2 tables), *Cramer's V* (for tables without the same number of columns and rows), *Goodman and Kruskal's Tau* (T), and *Lambda* (λ). These are discussed in more detail in advanced statistics books and online resources. For our purposes, the goal is to understand how to interpret correlation coefficients.

Consider again the results between political party and gender found in Table 7.1. Note that the Phi, Cramer's V, and Contingency coefficients are all fairly similar in Table 7.3, around 0.32 to 0.34 (rounding off), and all are statistically significant ($p < .001$). Chi-square told us that there is a relationship, and these statistics communicate that it's a moderate one in strength. Remember, chi-square does not let us know how strong or weak the relationship is, only that there is one.

Table 7.3 Example of Statistics for Nominal Variables, SPSS

SYMMETRIC MEASURES			
		Value	Approximate Significance
Nominal measures	Phi	0.339	.001
	Cramer's V	0.339	.001
	Contingency coefficient	0.321	.001
N of valid cases		151	

Lambda and Tau. Many researchers prefer to use coefficients that determine the *proportional reduction in error* (*PRE*). *Lambda* and *Goodman and Kruskal's Tau* are two statistics best suited for this when you have nominal variables. What PRE means is that the values of Lambda and Tau tell us approximately how many fewer errors we will make when predicting the outcome values of the dependent variable once we know the values of the independent variable. Be aware, though, that Lambda is based on modes, so it is possible to obtain a Lambda of 0. This simply means that a prediction of the dependent variable mode is not helped by knowing the modes of each category of the independent variable.

Because it is calculated using modes, the value of Lambda is a function of which variable is the dependent one. Note that several values are calculated, depending on whether the association you are assessing is symmetric or asymmetric. *Asymmetric* (or directional) measures assume that one of the variables is definitely dependent on the other; *symmetric* measures allow for the possibility of either one being dependent. When your hypothesis is one-directional, then asymmetric Lambda or Tau is more appropriate. If you are interested, for example, in the relationship between sex and race/ethnicity in your sample, then you would look at symmetric Lambda because neither variable occurs before the other in time or is being used to predict the other.

Because political party is the dependent variable, the value of Lambda is 0.000, the second one in Table 7.4. According to these statistics, there is no correlation between gender and political party, hence there is no significance level given. With most statistical output, focus primarily on two items: the value of the statistic and its significance level. The other information provided is useful for those with more advanced statistical training.

Lambda is calculated using the mode; the modal political party for men in this sample is Democrat, and the modal political party for women is also Democrat. Knowing respondents' genders does not improve or reduce our errors when predicting their political preference. Your best guess in either case is Democrat, regardless

Table 7.4 Example of Lambda and Tau Statistics for Nominal Variables, SPSS

DIRECTIONAL MEASURES

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Signif.
Nominal measures	Lambda	Symmetric political party	0.069	.046	1.424	.155
		Dependent	0.000	.000		
		Gender dependent	0.151	.098	1.424	.155
	Goodman and Kruskal's Tau	Political party				
		Dependent	0.057	.028		.000[c]
		Gender dependent	0.115	.053		.001[c]

[a] Not assuming the null hypothesis.
[b] Using the asymptotic standard error assuming the null hypothesis.
[c] Based on chi-square approximation.

of the gender. And that is your best guess in general because the mode for the overall sample is Democrat as well.

Here is how you interpret a PRE if there is some correlation coefficient greater than 0. Let's say you get a value for Lambda of 0.45 between sexual orientation and religious affiliation. This coefficient tells you that errors in predicting the religion of the respondents in your sample are reduced by a proportion of 0.45 (or 45 percent) when you know what their sexual orientations are. Or take another example: Imagine you and your friends go to someone who claims to read people's minds and will tell you what your occupations are. Suppose this person knew absolutely nothing about you; you hide behind a screen and don't even speak. The guesses will be all over the place because there are no hints or clues to help the mind reader. So you help out, and you and your friends disclose what your educational attainments are (high school grad, college grad, etc.). There will be fewer errors now since many college graduates tend to have different kinds of jobs than many others with only a high school diploma. By how much does the mind reader improve in predicting your occupations knowing your educational levels? The PRE tells you that. So, for nominal level data, Lambda and Tau provide you with this information: With a PRE correlation of 0.23, the mind reader would make 23 percent fewer errors in guessing

your jobs once educational levels are disclosed. Not only do you get a sense of the strength of the relationship and can compare the results to other relationship coefficients, you also know how well you will be reducing errors in predicting (PRE) and explaining the dependent variable.

Because the goals of research include explaining and predicting as much as we can about our outcome variables, it is important to calculate statistics that determine how well we are doing with our independent variables. Chi-square tells us whether our two variables are independent of each other, or to put it another way, whether they are associated. Once we know that they are related, coefficient measures like Tau and Lambda tell us how strongly the variables are associated and how much error reduction we have when we predict the dependent variable values knowing the independent variable values.

Measures of Association: Ordinal Variables

In addition to chi-square, several other statistics are appropriate when the categories of the variables are ordered. If you wish to assess the association of two ordinal variables, such statistics as Gamma (γ), Kendall's Tau-b and Tau-c, and Somers' d can be used to determine the strength of the association. These coefficients also range from 0.0 to 1.0 but, unlike the ones for nominal data, they can be negative or positive, depending on the order of the categories. Each of these statistics has sampling distributions that are used to determine the probability of obtaining that statistic by chance. If the probability of obtaining the value calculated for Gamma or Kendall's Tau-c by chance is less than the level you set (minimally .05), then you reject the null hypothesis and declare there is an association between the independent and dependent variables.

Gamma. A popular statistic to use is *Gamma*, which is based on a concept of evaluating pairs of responses and whether the respondents' relative order on both the independent and dependent variables is similar or dissimilar. Gamma is a symmetric statistic (only one coefficient is calculated because it doesn't matter which is independent or dependent) and a PRE measure. This means that the coefficient itself not only tells you the strength of the relationship but it also indicates the proportion of reducing error in predicting the dependent variable once you have information on the independent variable. See Table 7.5, which shows the results of a study that found a Gamma of 0.191 between two ordinal variables: political views (liberal, moderate, conservative) and attending religious services (several times a year or less versus once a month or more). Although statistically significant, this could be interpreted as a weak relationship in which 19 percent of the errors in predicting religious attendance are reduced by knowing respondents' political views.

Table 7.5 Example of Kendall's Tau and Gamma Statistics for Ordinal Variables, SPSS

SYMMETRIC MEASURES

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Signif.
Ordinal by ordinal	Kendall's Tau-b	0.111	0.026	4.298	.000
	Kendall's Tau-c	0.127	0.030	4.298	.000
	Gamma	0.191	0.044	4.298	.000
N of valid cases		1,316			

[a] Not assuming the null hypothesis.
[b] Using the asymptotic standard error assuming the null hypothesis.

Kendall's Tau. *Kendall's Tau-b* (not to be confused with Goodman and Kruskal's Tau for nominal level variables) uses different methods for calculation than Gamma and assumes there are the same number of rows and columns in the table. *Kendall's Tau-c* can be used for tables of data in which the numbers of rows and columns are not equal. In this example, because there are three categories for political views and two for religious attendance, Tau-c could be used. These are symmetric statistics, and the same coefficient of 0.127 would result if we decided political views depended on religious attendance. But Kendall's statistics are not PRE measures. The coefficient calculated can only be interpreted in terms of strength (weak to strong) and direction (positive or negative/inverse relationship). In this case, it is positive but weak.

Somers' d. Similar to Gamma in the way it is calculated (that is, based on comparing the similarity or dissimilarity of ranking on pairs of responses on the two variables), *Somers' d* differs in that it can be an asymmetric statistic where two values are calculated, depending on which is the dependent variable. However, it does not have a PRE interpretation.

In Table 7.6, the correlation between political views (in order: liberal, moderate, conservative) and attending religious services (several times a year or less versus once a month or more) is 0.110 when you assume it is symmetrical and one variable is not predicting the other. If you feel that political views precede religious attendance, then the value of Somers' d is 0.096. In either case, it is statistically significant ($p < .001$) but a weak positive coefficient. This means there is some relationship between those who are conservative (a high value on the political views measure) and attending services once a month or more (a high value on the religious services measure), or conversely, those who are more liberal in this sample seem to attend religious services

Table 7.6 Example of Somers' d Statistic for Ordinal Variables, SPSS

DIRECTIONAL MEASURES

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Signif.
Ordinal by ordinal	Somers' d Symmetric	0.110	0.025	4.298	.000
	Religious services attendance dependent	0.096	0.022	4.298	.000
	Political views dependent	0.128	0.030	4.298	.000

[a] Not assuming the null hypothesis.
[b] Using the asymptotic standard error assuming the null hypothesis.

less frequently than those who are conservative. It is statistically significant at such a low level of strength because it is from a large sample, in this case, a sample of over 1,300 respondents.

The differences among all these ordinal statistics are the way they deal differently with pair rankings that are tied, something discussed in more detail in advanced statistics books or online resources. Yet notice how similar they are in terms of the actual coefficient values. However, if you want a statistic that not only provides a measure of the strength of the association but also indicates how much error reduction occurs, then Gamma is ideal for ordinal measures.

Spearman's Rho. Sometimes both the independent and dependent variables consist of rank ordered numbers. For example, you want to compare the rank order of people in terms of their grades on the first research methods exam with their rank order on the last exam taken. Student A might be ranked third on the first test and fifth on the last test, student B might be twelfth on the first test and second on the last test, and so on for each of the 100 students in your study. Note that you are comparing not actual grades (which are interval/ratio measures) but rather their rank order number. In such cases, *Spearman's rank order correlation coefficient* or *rho* (ρ) provides a measure of association.

Spearman's rho is a symmetrical statistic that results in a coefficient between 0.0 and 1.0 to indicate the strength of the relationship and a plus or minus to show the direction of the relationship. In addition, it's a PRE measure when you square the value of rho. For example, if the correlation coefficient is 0.65 between the first set of grades and the last set, then you can conclude that there is a strong positive relationship between

the two and that error in predicting rank order on the last exam is reduced by around 42 percent (0.65 squared). If for some reason the coefficient were −0.65, then you would conclude that those who rank higher on the first exam tend to rank lower on the last exam, and those who rank lower on the first rank higher on the last.

Measures of Association: Interval/Ratio Variables

Ranking scores on a test, as in the previous example, loses information. The person with the best grade could be only 1 point higher than the next person or 20 points higher. Ordinal measures do not give you enough information like the distance between the ranks. Certainly, if you have interval/ratio measures, then make use of them, because they represent the best level of measurement for the more sophisticated statistical analyses. There are several ways you can evaluate a relationship between interval/ratio variables.

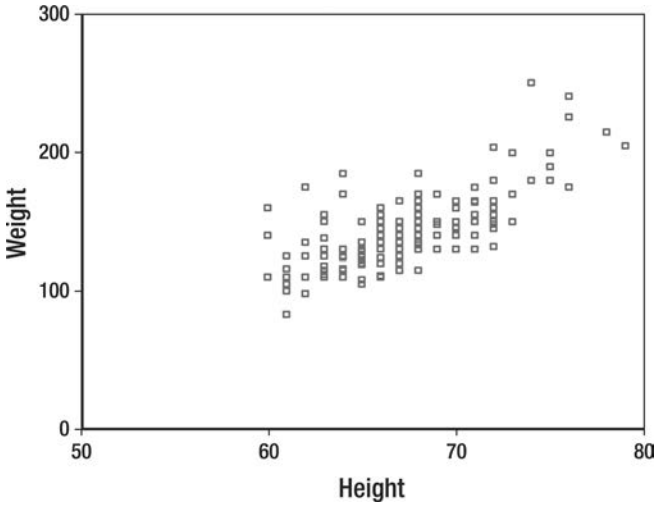
Scatterplots. Although you are not likely to put interval/ratio variables into cross-tabulations when there are a large number of values for each variable, you can visually assess the relationship with a graph. Most common are *scatterplots*. As with frequency curves for univariate data, scatterplots assume interval/ratio measures. The horizontal x-axis is usually for the values of the independent variable, and the vertical y-axis for the units of the dependent variable. A mark is placed at the point where each surveyed person's responses to the two variables intersect, not unlike the idea of a cell in a contingency table for nominal and ordinal variables. When all the responses are plotted, a pattern emerges.

When the points are scattered everywhere in the graph with no apparent pattern, it is signifying no relationship, that is, an association close to 0.0 coefficient. When the points tend to fall along an almost straight line, then there is a perfect relationship close to 1.0. If it looks like an uphill line, then it's a positive linear relationship; if it looks like a downhill line, it's a negative/inverse linear one. When the points tend to form a curve, then it's called a curvilinear relationship.

Take, for example, the relationship between height and weight. Usually, taller people weigh more, and shorter people weigh less. But as we all know, this is not a perfect guide; there are many short, heavy people and tall, thin people. We would expect then a scatterplot that looks like an uphill slope, but not all the points would fall in a perfect straight line.

As you can see from the scatterplot in Figure 7.2, there is a tendency for those who are shorter (the x-axis marks the height in inches) to be lighter in weight (the y-axis marks the weight in pounds). Each point on the chart represents one person's height and weight. If you were to draw an exact straight line from any point to the x-axis, you would find out the height for that respondent, and a line from the same point to the

Figure 7.2 Example of a Scatterplot, SPSS



y-axis would indicate the weight. Notice that there are some outliers, that is, respondents who fall beyond an imaginary straight line that could be drawn through the middle of the points. For example, there appears to be three people around 60 inches tall but one of the points is closer to 175 pounds or so on the y-axis. That person's data falls outside the general pattern.

Pearson r. Although scatterplots are not as precise as a table of numbers, they do provide a rough visual indication of the kind of relationship that exists between two variables. However, for a more specific assessment of the magnitude (weak to strong) and direction of the relationship (positive or negative/inverse), *Pearson's product-moment correlation coefficient*, a symmetric statistic known simply as *Pearson r*, is calculated. Similar to the other coefficients discussed, it has two components: a number ranging anywhere from 0.0 to 1.0 to indicate strength, and a plus or minus sign to show direction. Like Gamma, Lambda, and Goodman and Kruskal's Tau, Pearson r is also a PRE measure, but like Spearman rho, it must be squared to establish the proportion of error reduction.

Pearson r is one of the most used statistics in social science research and central to path analysis, linear regression analysis, and other statistical methods. Its calculation is based on means, standard deviations, and z-scores. In order to compare variables with two different units of measurement, such as height and weight, respondents' answers are transformed into z-scores (as described in Chapter 6). What Pearson r does is measure how much change in the z-scores of one variable is related to change in the z-scores of the other variable. Is the variation from respondent A's z-score for height to that of respondent B's similar to the variation from respondent A's z-score

for weight to respondent B's, and so on throughout the sample of all respondents? Every change in an independent variable z-score (a standard deviation unit) is related to a change in a dependent variable z-score. That is why it is called a correlation. If those changes vary at the same rate and with the same incremental units, then there is a perfect 1.0 correlation coefficient that looks like a straight line in a scatterplot, either going up if positive (/) or going down if it's an inverse relationship (\).

The data correlating height and weight result in a Pearson r of 0.714 (Table 7.7). For every change of one standard z-score unit in height (independent variable) there is a 0.714 standard unit change on the outcome or dependent variable (weight) in the same direction. Those who are taller tend to weigh more, and those who are shorter tend to weigh less. If it were an inverse relationship of −0.714, then we would conclude that taller people weigh less and shorter people weigh more. The correlation does not tell you about any one person, so it is not accurate to say that any one short person weighs less. Making inferences about any one individual in the sample from group data is called the *ecological fallacy*. These are aggregated data and are only interpretable in terms of the entire sample, not in terms of one person at a time. Because the goal is to predict variation across the sample on the dependent variable, you can also say that you have reduced errors in predicting the weights of all respondents by 51 percent (r^2), by knowing information about their heights.

The properties of this statistic also allow you to say it another way: You have determined 51 percent of the variation in weight (the dependent variable) through your knowledge of height (the independent variable). This is why r^2 is called the *coefficient of determination*. In this particular example, of all the variation that exists in your sample for weight (after all, not everyone is the same weight), you account for 51 percent of it with the respondents' heights. Height is only one determinant (explanation or predictor), albeit a major one, of the outcome, namely, weight. The remaining

Table 7.7 Example of Pearson r Statistic for Interval/Ratio Variables, SPSS

CORRELATIONS

		Height	Weight
Height	Pearson correlation	1	0.714*
	Significance (two-tailed)		.000
	N	151	149
Weight	Pearson correlation	0.714*	1
	Significance (two-tailed)	.000	
	N	149	150

*Correlation is significant at the .01 level.

49 percent of the variation (sometimes referred to as the *residual*) must be explained or predicted by other reasons such as genetics, food habits, and amount of exercise.

As with the other statistics, Pearson r has a sampling distribution, and the data analysis software program can use it to calculate what the probability of obtaining a particular coefficient is by chance alone. If the probability is less than .05, we reject the null hypothesis and declare there is a statistically significant relationship between the independent and dependent variables. However, when sample sizes are large, small coefficients tend to be statistically significant, and for this reason, many researchers look primarily to the strength of the relationship and not as much to its significance level.

Eta. Pearson r also requires that the relationship between the two variables is linear; that is, it is not a curve. As one variable increases or decreases, so too does the other. An example of a nonlinear relationship might be age and muscle strength. As you grow older, you get stronger, but then after a certain age, your muscles get weaker. Strength does not continue to grow linearly as you get older. Pearson r would not be an adequate measure of the relationship between age and physical strength. It would be more appropriate to use *eta* (η), which is designed for nonlinear associations. It requires an interval/ratio dependent variable, and the independent variable should be in nominal or ordinal categories. When squared, eta is a PRE measure and can be interpreted similarly to r^2. So if you find an eta of 0.50 between age and muscle strength, then you can say that you can reduce your errors in predicting physical strength by 25 percent by knowing different categories of age, such as younger people are stronger and older people are weaker.

The Purposes of Measuring Relationships

Remember that the main goals of doing research are to describe, explain, or predict. The objective is to account for the reasons why the dependent variable varies among the respondents in your study, to predict future occurrences, or simply to describe any relationships among your variables. Measures of association are a good way of evaluating these relationships, to see if they exist at all and, if so, how strong they are and in what direction they go. It also becomes necessary sometimes to elaborate on the original relationship, to control for alternative explanations, and to test for spurious relationships. Chapter 9 discusses how to elaborate your findings by introducing a third or control variable into the analyses.

When we analyze relationships and generate PRE information, we answer some of the research questions or hypotheses we began with, thereby contributing to theory building, to the assessment and evaluation of programs, and to descriptions of some social phenomena. However, sometimes we are interested in describing, explaining, or predicting differences among various categories of people and not just uncovering associations between variables. Ways of analyzing data for differences are discussed in the next chapter.

REVIEW: WHAT DO THESE KEY TERMS MEAN?

Bivariate data analysis
Cells
Chi-square
Correlation coefficient
Coefficient of determination
Contingency tables
Cross-tabulations (crosstabs)
Degrees of freedom
Ecological fallacy
Eta
Gamma
Goodman and Kruskal's Tau
Kendall's Tau-b and Tau-c
Lambda
Marginals
Pearson r
PRE
Residual
Rows and columns
Scatterplots
Somers' d
Spearman's rho
Symmetric and asymmetric

TEST YOURSELF

Here are several hypotheses: First, identify independent and dependent variables and their levels of measurement. Then, say which statistic would be most useful to test the association between them.

1. There is no relationship between high school grades (GPA) and college grades (GPA).

	Which variable?	Level of measurement?	Which statistic to use?
Independent variable			
Dependent variable			

2. There is no relationship between type of car owned and region of the country (rural or urban).

	Which variable?	Level of measurement?	Which statistic to use?
Independent variable			
Dependent variable			

3. There is no relationship between the football rankings of a set of universities this year compared with their rankings last year.

	Which variable?	Level of measurement?	Which statistic to use?
Independent variable			
Dependent variable			

4. There is no relationship between age (measured as "under 18," "18 to 25," "26 to 33," "34 to 41," and "42 and older") and number of times using Twitter per day (measured as "none," "1 to 5 times," "6 to 10 times," "11 or more times").

	Which variable?	Level of measurement?	Which statistic to use?
Independent variable			
Dependent variable			

INTERPRET: WHAT DO THESE REAL EXAMPLES TELL US?

1. Table 7.8 shows some SPSS output from the 2014 General Social Survey.
   a. Which is the strongest correlation in this matrix? Note that the table is symmetrical because the correlation of "Hours per Day Watching TV" with "Highest Year of School Completed" is the same as "Highest Year of School Completed" with "Hours per Day Watching TV" and that correlations of variables with themselves always equal a perfect 1.0 correlation.

Table 7.8 Pearson r Correlations

CORRELATIONS

		Highest year of school completed	*Hours per day watching TV*	*Age of respondent*
Highest year of school completed	Pearson correlation	1	−0.206**	−0.014
	Significance (two-tailed)		.000	.479
	N	2,537	1,668	2,528
Hours per day watching TV	Pearson correlation	−0.206**	1	0.138**
	Significance (two-tailed)	.000		.000
	N	1,668	1,669	1,666
Age of respondent	Pearson correlation	−0.014	0.138**	1
	Significance (two-tailed)	.027	.000	2,529
	N	2,528	1,666	

**Correlation is significant at the .01 level (two-tailed).

b. How would you put into words the relationship between education and television viewing? What does the minus sign tell you?
c. The correlation between age and years of schooling completed is statistically significant. Yet over 2,500 respondents answered these questions. How would you interpret this significance level in comparison to what seems to be a weak correlation?

2. Sotoudeh et al. (2017) surveyed thousands of Facebook users in Muslim-majority countries to evaluate the relationships between romance in public spaces and in cyberspaces. For example, the researchers wanted to know if single people who hold hands in public with members of the opposite sex use the Internet to arrange dates on online dating sites. This is what they found:

	Not held hands	Held hands	Total
Not arrange dates	90% (3,233)	78% (1,623)	3,596
Arrange dates online	10% (363)	22% (458)	2,081
Total	100% (4,856)	100% (821)	5,677

Chi-square test 150.29 p <.01

a. State the null hypothesis being tested in this table.
b. Put into words what is going on. How do you read the table? For example, 78 percent of what is what, and so on.
c. What does the chi-square tell you? What does the p value mean?
d. What do you conclude?
e. Which other statistics could you use to assess the strength of the relationship?

CONSULT: WHAT COULD BE DONE?

You are hired to do a marketing research analysis for the local newspaper. The publishers want to find out which kinds of people use the Internet to read the online version of the paper instead of buying the print version, how often, and which sections and features are most read and liked.

1. Preliminary data analysis suggests that men and women read different sections of the online paper. What would you do to see if this is statistically significant?
2. Frequency of reading the online version of the paper seems to vary on the basis of such characteristics as income, educational level, age, and size of family. How would you statistically test each of these possible relationships (income and frequency

of reading the paper, education and frequency, age and frequency, and size and frequency)?
3. What other relationships would you recommend be studied to provide the newspaper owners more information about readership?

DECIDE: WHAT DO YOU DO NEXT?

For your study on how people develop and maintain diverse friendships, especially on social media, respond to the following items:

1. Review your bivariate hypotheses (or write a few new ones) and list the variables.
2. Describe the statistics you would use to evaluate the relationships and why you would use these.
3. If you have actual data, begin to analyze the hypotheses that are best suited for statistical measures of association and correlation.