Chapter thirteen

Measures of Association for Nominal and Ordinal Variables

Measuring the strength of a relationship

Measures of association for nominal variables

What Do Nominal Measures of Association Describe?

When are These Measures Used?

What is the Test of Statistical Significance for These Measures?

Measures of association for ordinal variables

What Do Ordinal Measures of Association Describe?

When are These Measures Used?

What is the Test of Statistical Significance for These Measures?

D. Weisburd and C. Britt, *Statistics in Criminal Justice*, DOI 10.1007/978-1-4614-9170-5_13,

Chapter 12 introduced eta (η) and the more general concept of measures of association. Eta is a descriptive statistic that allows us to define how strongly the categorical variable or sample in an analysis of variance is related to the interval-level variable or trait we examined across the samples. But there are many other useful measures of association that allow us to define relationships among variables. Over the next few chapters, we will focus on some of these that are particularly useful in studying criminal justice. We will still be concerned with statistical significance in these chapters, but we will examine not only whether a measure is statistically significant but also how strong the relationship is.

In this chapter, our focus is on nominal- and ordinal-level measures of association. We begin with a discussion of why it is important to distinguish between statistical significance and strength of association. While statistical significance can tell us whether we can make reliable statements about differences in a population from observations made from samples, it does not define the size of the relationship observed. It is important to define the strength of the relationship between variables being examined because that puts us in a better position to decide whether results that are statistically significant are also substantively important.

Distinguishing Statistical Significance and Strength of Relationship: The Example of the Chi-Square Statistic

In Chapter 9, we explored the chi-square statistic as a way to determine whether there was a statistically significant relationship between two nominal-level variables. The chi-square statistic is useful as a way of testing for such a relationship, but it is not meant to provide a measure of the strength of the relationship between the variables. It is tempting to look at the value of the chi-square statistic and the observed significance level associated with a particular chi-square value and infer from these statistics the strength of the relationship between the two variables. If we follow such an approach, however, we run the risk of an interpretive error.

Table 13.1 Observed Frequencies (f_o) and Expected Frequencies (f_e) for Two Outcomes of an Experimental Condition with 200 Cases

EXPERIMENTAL CONDITION	OUTCOME		
	Failure	Success	Total
Treatment	$f_o = 40$ $f_e = 50$	$f_o = 60$ $f_e = 50$	100
Control	$f_o = 60$ $f_e = 50$	$f_o = 40$ $f_e = 50$	100
Total	100	100	200

The problem with using the chi-square statistic—or outcomes of other tests of statistical significance—in this way is that the size of the test statistic is influenced not only by the nature of the relationship observed but also by the number of cases in the samples examined. As we have noted a number of times in the text, this makes good sense. Larger samples, all else being equal, are likely to be more trustworthy. Just as we feel more confident in drawing inferences from a sample of 10 or 20 coin tosses than from a sample of 2 or 3 tosses, our confidence in making a decision about the null hypothesis grows as the sizes of the samples examined using a chi-square statistic increase.

The following example will help to illustrate this problem. Suppose we have a sample of 200 cases that cross-tabulate experimental condition with an outcome measure, as shown in Table 13.1. We see that 60% of those in the treatment group have an outcome classified as a success, while only 40% of those in the control group have an outcome classified as a success. Our calculated value of chi-square for these data is 8.00 with df = 1, which has an observed significance level less than 0.01 (see Appendix 2). See Table 13.2 for detailed calculations for obtaining the chi-square statistic.

Table 13.2 Calculations for Obtaining Chi-Square Statistic for the Example in Table 13.1

EXPERIMENTAL CONDITION	OUTCOME	f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
Treatment	Failure	40	50	−10	100	2
Treatment	Success	60	50	10	100	2
Control	Failure	60	50	10	100	2
Control	Success	40	50	−10	100	2
						$\Sigma = 8.0$

Without changing the proportional distribution of cases for this example—keeping success at 60% for the treatment group and 40% for the control group—suppose we multiply the number of cases by 10. We now have 2,000 total observations, as shown in Table 13.3, but the relationship between experimental condition and outcome is the same. Our calculated chi-square statistic, however, now has a value of 80.00 (see Table 13.4) with df = 1, and the observed significance level is less than 0.0001. So, simply by increasing the size of the sample, we increase the value of chi-square and decrease the corresponding observed significance level.

This feature of the chi-square statistic applies to all tests of statistical significance. Irrespective of the observed relationship between measures, as the sample size increases, the observed significance level associated with that relationship will also increase. This simple rule regarding the relationship between statistical significance and sample size will be examined in more detail in the discussion of statistical power in Chapter 21. The rule does not raise any new questions regarding the meaning of statistical significance. It simply reminds us that, all else being equal, we can be more confident in making statistical inferences from larger samples. It also emphasizes the importance of distinguishing between statistical significance and the size or strength of a relationship between variables.

To allow researchers to define the strength of a relationship among nominal-level or ordinal-level variables, statisticians have developed a variety of measures of association. Some of these measures are based on the value of the chi-square statistic; others are based on unique transformations of the counts or distributions of cases within a table. All the measures of association that we discuss share a standardized scale: A value of 0 is interpreted as no relationship, and a value of 1.0 (or, in the case of ordinal scales, +1 or −1) is interpreted as a perfect relationship between the two variables. The discussion that follows describes some of the more frequently used measures of association for nominal and ordinal variables.

Table 13.3 Observed Frequencies (f_o) and Expected Frequencies (f_e) for Two Outcomes of an Experimental Condition with 2,000 Cases

EXPERIMENTAL CONDITION	OUTCOME		Total
	Failure	Success	
Treatment	$f_o = 400$ $f_e = 500$	$f_o = 600$ $f_e = 500$	1,000
Control	$f_o = 600$ $f_e = 500$	$f_o = 400$ $f_e = 500$	1,000
Total	1,000	1,000	2,000

Table 13.4 Calculations for Obtaining Chi-Square Statistic for the Example in Table 13.3

EXPERIMENTAL CONDITION	OUTCOME	f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
Treatment	Failure	400	500	−100	10,000	20
Treatment	Success	600	500	100	10,000	20
Control	Failure	600	500	100	10,000	20
Control	Success	400	500	−100	10,000	20
						$\Sigma = 80.0$

Measures of Association for Nominal Variables

Measures of Association Based on the Chi-Square Statistic

The preceding example illustrated how the chi-square statistic is affected by sample size. With a 2 × 2 table (i.e., two rows and two columns), one straightforward way of measuring the strength of a relationship between two variables that adjusts for the influence of sample size is to transform the value of the chi-square statistic by adjusting for the total number of observations. One measure of association that does this is **phi (ϕ).** Phi is obtained simply by dividing the chi-square statistic by the total number of observations (N) and taking the square root of this value (see Equation 13.1).

$$\phi = \sqrt{\frac{\chi^2}{N}} \qquad \textbf{Equation 13.1}$$

Phi will have a value of 0 if the value of the chi-square statistic is 0 and there is no relationship between the two variables. Phi will have a value of 1 if the chi-square statistic takes on a value equal to the sample size, which can occur only when there is a perfect relationship between two categorical variables. It is important to note that phi is appropriate only for analyses that use a 2 × 2 table. If the number of rows or columns exceeds two, then it is possible for phi to take on values greater than 1.0, eliminating the possibility of any kind of meaningful interpretation.

Consider the two chi-square statistics that we calculated above for the data in Tables 13.1 and 13.3: 8.00 and 80.00, respectively. If we insert these values for chi-square and the sample size, we find that the value of phi for *both* tables is 0.20.

Working It Out

$$\phi = \sqrt{\frac{8.00}{200}} = 0.20 \qquad \text{and} \qquad \phi = \sqrt{\frac{80.00}{2{,}000}} = 0.20$$

We now have a measure of association that is not influenced by sample size. For both of our examples, in which the proportion of cases in each group was similar, we have the same phi statistic. However, is the relationship large or small? As noted in Chapter 12, defining "large" and "small" is a matter of judgment and not statistics. In judging the importance of a result, researchers can compare it with other findings from prior studies. Or they may examine the importance of the policy implications that could be drawn from the result. For example, a very small change in rates of heart attacks in the population could save many lives, and thus a small relationship may still be important. According to a standard measure of effect size suggested by Jacob Cohen, a phi of 0.10 is considered to indicate a small relationship, one of 0.30 a medium relationship, and one of 0.50 a large relationship.[1]

Our examples suggest why we might be misled if we used the chi-square statistic and its corresponding significance level as an indicator of the strength of the relationship between two variables. If we had tried to infer the strength of the relationship between experimental condition and outcome from the value of the chi-square statistic, we would have been tempted to conclude that Table 13.3 showed a stronger relationship than Table 13.1. However, once we take into account the size of the sample, we see that the two tables reflect the same relationship between the two variables. The data in Table 13.3 lead to a higher observed significance level because the samples examined are larger. However, the strength of the relationship observed in the two tables is the same.

For tables with more than two rows or two columns, we cannot use phi. Instead, we use a measure of association known as **Cramer's *V,*** which is also based on the value of the chi-square statistic but makes an adjustment for the number of categories in each variable. Equation 13.2 presents the formula for calculating Cramer's *V.*

$$V = \sqrt{\frac{\chi^2}{N \times \min(r - 1,\ c - 1)}} \qquad \textbf{Equation 13.2}$$

In Equation 13.2, the chi-square statistic (χ^2) is divided by the product of the total number of observations (N), and the smaller of two numbers,

[1]See Jacob Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Hillsdale, NJ: Lawrence Erlbaum, 1988), pp. 215–271.

$r - 1$ or $c - 1$ (i.e., the minimum of these two values), where r is the number of rows in the table and c is the number of columns. For example, if we had a table with two rows and three columns, we would have $r - 1 = 2 - 1 = 1$ and $c - 1 = 3 - 1 = 2$. The value for $r - 1$ is the smaller of these two numbers, so we would use that value (1) for $\min(r - 1, c - 1)$ in the denominator of the formula. If we were working with a larger table with, say, five rows and four columns, we would have $r - 1 = 5 - 1 = 4$ and $c - 1 = 4 - 1 = 3$. Since 3 is less than 4, we would use the value 3 for $\min(r - 1, c - 1)$ in the denominator.

Let's consider an example. Table 13.5 reproduces the data from Table 9.9 on cell-block assignment and race of prisoner. Recall from Chapter 9 that the chi-square statistic for this cross-tabulation was 88.2895, and with df = 6, the observed significance level was less than 0.001. We can use the data in this table to illustrate the calculation of V. The table has seven rows ($r = 7$) and two columns ($c = 2$), meaning that $r - 1 = 7 - 1 = 6$ and $c - 1 = 2 - 1 = 1$. The smaller of these two values is 1, which we substitute for $\min(r - 1, c - 1)$ in the denominator of the formula for V. After inserting the other values into Equation 13.2, we find that $V = 0.2708$.

Working It Out

$$V = \sqrt{\frac{\chi^2}{N \times \min(r - 1, c - 1)}}$$

$$= \sqrt{\frac{88.2895}{(1{,}204)(1)}}$$

$$= 0.2708$$

Table 13.5 Assignment of Non-Hispanic White and Nonwhite Prisoners in Seven Prison Cell Blocks

CELL BLOCK	NON-HISPANIC WHITES	NONWHITES	ROW TOTAL
C	48	208	256
D	17	37	54
E	28	84	112
F	32	79	111
G	37	266	303
H	34	22	56
I	44	268	312
Column total	240	964	1,204

The value of Cramer's V may be interpreted in the same way as that of phi. Accordingly, a value for V of 0.2708 is suggestive of a moderate relationship between cell-block assignment and race of prisoner.

Proportional Reduction in Error Measures: Tau and Lambda

Some measures of association that are appropriate for nominal-level variables are based on the idea of **proportional reduction in error,** or **PRE.** Such measures indicate how much knowledge of one variable helps to reduce the error we make in defining the values of a second variable. If we make about the same number of errors when we know the value of the first variable as when we don't, then we can conclude that the PRE is low and the variables are not strongly related. However, if knowledge of one variable helps us to develop much better predictions of the second variable, then we have a high PRE and the variables may be assumed to be strongly related.

Two of the more common measures of association between nominal variables, **Goodman and Kruskal's tau (τ)** and **lambda (λ)** are both PRE measures. Both of these measures require that we identify at the outset which variable is the **dependent variable** and which variable is the **independent variable.** A dependent variable is an outcome variable—it represents the phenomenon that we are interested in explaining. It is "dependent" on other variables, meaning that it is influenced—or we expect it to be influenced—by other variables. Any variable that affects, or influences, the dependent variable is referred to as an independent variable. The values of Goodman and Kruskal's tau (τ) and lambda (λ) will generally differ depending on which variable is identified as the dependent variable and which as the independent variable.

For most research projects, a body of prior research and/or theory will indicate which variables are dependent and which are independent. For example, for the study in Table 13.1, the independent variable is the experimental condition: the treatment or the control group. Whether the person participated in the treatment or the control group is generally theorized to influence outcome success or failure, which is the dependent variable. In other words, the experiment tests whether success or failure is due, at least in part, to participation in some kind of treatment.

PRE measures of association, such as tau and lambda, require the use of two decision rules. The first decision rule—the *naive decision rule*—involves making guesses about the value of the dependent variable without using any information about the independent variable. The second decision rule—the *informed decision rule*—involves using information about how the cases are distributed within levels or categories of the independent variable. The question becomes "Can we make better predictions about the value of the dependent variable by using information about the independent variable?" Will the informed decision rule provide

Table 13.6 Hypothetical Distribution of 200 Cases for Two Nominal Variables

(a) PRE Measure of Association = 0.0

	VARIABLE 2		
VARIABLE 1	**Category 1**	**Category 2**	**Row total**
Category 1	50	50	100
Category 2	50	50	100
Column total	100	100	200

(b) PRE Measure of Association = 1.0

	VARIABLE 2		
VARIABLE 1	**Category 1**	**Category 2**	**Row total**
Category 1	0	100	100
Category 2	100	0	100
Column total	100	100	200

better predictions than the naive decision rule? PRE measures of association have a value of 0 when there is no relationship between the two variables and a value of 1 when there is a perfect relationship between the two variables. Table 13.6 presents two hypothetical distributions illustrating PRE measures showing no relationship (part a) and a perfect relationship (part b). In part a, we see that knowledge of one variable does not help us make predictions about the second variable, since the cases are evenly distributed across all possible cells of the table (e.g., of the 100 cases in Category 1 of Variable 1, exactly 50 cases each fall into Category 1 and 2 of Variable 2). In the perfect relationship shown in part b, knowledge of one variable determines, without error, the value of the second variable (e.g., all cases in Category 1 of Variable 1 fall into Category 2 of Variable 2).

A key advantage to PRE measures of association is the interpretation of values between 0 and 1. Any value greater than 0 may be interpreted as a proportionate reduction in error achieved by using information on the independent variable. Alternatively, we can multiply the PRE measure by 100 and interpret the value as the percent reduction in errors. For example, a PRE measure of 0.50 indicates a percent reduction in prediction errors of 50% when information about the independent variable is used.

For an illustration of the calculation of tau and lambda, consider the data presented in Table 13.7. These data come from responses to a survey by adult residents of the state of Illinois.[2] Respondents who reported

[2]For a description of the study, see Chester L. Britt, "Health Consequences of Criminal Victimization," *International Review of Victimology,* 8 (2001): 63–73.

Table 13.7 Data on Victim-Offender Relationship and Location of Assault

VICTIM-OFFENDER RELATIONSHIP	LOCATION OF ASSAULT				
	Home	Neighborhood	Work	Someplace Else	Row total
Stranger	10	49	18	89	166
Acquaintance/friend	21	22	7	46	96
Partner	77	5	3	19	104
Relative	31	1	2	10	44
Column total	139	77	30	164	410

that they had experienced an assault were asked a series of follow-up questions about the most recent event. Two of these questions addressed the relationship between the victim and the offender and the location of the assault. Here we have classified the victim-offender relationship into four categories: stranger, acquaintance/friend, partner (includes spouse and boyfriend or girlfriend), and relative. Location of the assault is also classified into four categories: home, neighborhood, work, and someplace else. For this analysis, we assume that the victim-offender relationship is the independent variable and the location of the assault is the dependent variable. Our research question is "What is the strength of the relationship between victim-offender relationship and location of an assault?"

Goodman and Kruskal's tau uses information about the marginal distributions of the two variables to test whether knowledge of the independent variable reduces prediction errors for the dependent variable. The first step in computing this statistic is to ask how many errors we would expect to make, on average, if we did not have knowledge about the victim-offender relationship. This is our naive decision rule, where we are effectively trying to guess what category of the dependent variable an observation might belong to, without using any information about the independent variable. For our example, we begin by looking at the column totals in Table 13.7, which reflect the categories of the dependent variable. Of the 410 assaults, we see that 139 occurred in the home, 77 in the neighborhood, 30 at work, and 164 someplace else. We use these column totals to help us determine the average number of errors we would expect to make if we assigned cases without any information about the victim-offender relationship.

Let's begin with assaults in the home. Of the 410 total assaults, 139 belong in the assaulted-in-the-home category, while 271 do not belong in this category (i.e., the assault occurred elsewhere). Proportionally, 0.6610 (271 of 410) of the cases do not belong in the assaulted-in-the-home category. If we randomly assigned 139 of the 410 cases to the assaulted-in-the-home category, we would expect 0.6610 of these 139

cases to be assigned incorrectly. To obtain the number of cases assigned incorrectly—the number of prediction errors—we multiply the proportion of cases not in the category by the number of cases assigned to that category. For assaulted in the home, this is (0.6610 × 139) = 92. The value 92 represents the number of prediction errors we would expect to make, on average, in assigning cases to the assaulted-in-the-home category without any knowledge of the victim-offender relationship.[3]

Turning to assaults in the neighborhood, we see that 77 cases belong in this category, and the remaining 333 do not belong in this category. As a proportion, 0.8122 of the cases (333 of 410) do not belong in the assaulted-in-the-neighborhood category. This means that we would expect to make 0.8122 × 77 = 63 prediction errors, on average, in assigning cases to this category without any knowledge of the victim-offender relationship. For assaults at work, 30 cases belong in this category and 380 do not, meaning that we would expect to make (380/410) × 30 = 28 prediction errors, on average, in assigning cases to the assaults-at-work category without any information about the victim-offender relationship. There are 164 cases that belong to the assaulted-someplace-else category, meaning that 246 cases do not belong in this category. We would expect to make (246/410) × 164 = 98 prediction errors, on average, in assigning cases to this category without any information about the victim-offender relationship. To determine the total number of prediction errors we would make without any knowledge of the victim-offender relationship, we add these four values together: 92 + 63 + 28 + 98 = 281 total prediction errors.

If we then use information about the victim-offender relationship—whether the victim and offender were strangers, acquaintances/friends, partners, or relatives—we can test whether this information improves our ability to predict the location of the assault. This reflects the use of our informed decision rule: Does our assignment of cases to categories of the dependent variable improve when we use information about the category of the independent variable? In other words, does knowing the category of the independent variable (victim-offender relationship) reduce the number of prediction errors we make about the category of the dependent variable (location of assault)? To the extent that the independent variable has a relationship with the dependent variable, the number of prediction errors should decrease.

The logic behind calculating the prediction errors is the same as before, except that we focus on the row totals in the table, rather than the total number of cases in each category of the dependent variable. We start with the first category of the independent variable (i.e., the first row of Table 13.7) and note that 166 cases involved offenders who were strangers to

[3]For all calculations of prediction errors, we have rounded the result to the nearest integer.

the victim. In a process similar to our earlier analysis, we begin by noting the placement of cases within this row: 10 assaults occurred at home, 49 in the neighborhood, 18 at work, and 89 someplace else. Starting with the assignment of cases to assaulted-in-the-home, we note that 10 cases belong in this category and 156 do not belong in this category. As a proportion, 0.9398 of the cases (156 of 166) do not belong in the assaulted-in-the-home category when the offender is a stranger. Thus, if we randomly assigned 10 of the 166 cases in this row to the assaulted-in-the-home category, we would expect to make $0.9398 \times 10 = 9$ prediction errors, on average. Turning to the assaulted-in-the-neighborhood category, we note that 49 cases belong in this category and 117 do not belong in this category, which means that we would expect to make $(117/166) \times 49 = 35$ prediction errors. For the assaulted-at-work category, we would expect to make $(148/166) \times 18 = 16$ prediction errors, and for the assaulted-someplace-else category, we would expect to make $(77/166) \times 89 = 41$ prediction errors. The total number of prediction errors in assigning cases involving offenders who were strangers is 101 (that is, $9 + 35 + 16 + 41$).

To determine the prediction errors for each of the remaining categories of the independent variable (assaults involving offenders who were acquaintances/friends, partners, or relatives), we use the same approach with the three remaining rows of Table 13.7. Table 13.8 presents all the calculations of prediction errors necessary for obtaining tau.

We obtain the total number of prediction errors made using information about the victim-offender relationship by summing the errors across each category of relationship. For cases involving an offender who was a stranger, we would expect to make 101 prediction errors; for cases involving an acquaintance or friend, we would expect to make 63 prediction errors; for cases involving partners, 44 prediction errors; and for cases involving a relative, 20 prediction errors (see the bottom row of Table 13.8). Altogether, we would expect to make 228 (that is, $101 + 63 +$

Table 13.8 Calculations of Prediction Errors for Obtaining Tau for a Relationship Between Victim-Offender Relationship and Location of Assault

LOCATION OF ASSAULT	PREDICTION ERRORS: No Knowledge of Victim-Offender Relationship	PREDICTION ERRORS: Offender Was a Stranger	PREDICTION ERRORS: Offender Was an Acquaintance or a Friend	PREDICTION ERRORS: Offender Was a Partner	PREDICTION ERRORS: Offender Was a Relative
Home	139(271/410) = 92	10(156/166) = 9	21(75/96) = 16	77(27/104) = 20	31(13/44) = 9
Neighborhood	77(333/410) = 63	49(117/166) = 35	22(74/96) = 17	5(99/104) = 5	1(43/44) = 1
Work	30(380/410) = 28	18(148/166) = 16	7(89/96) = 6	3(101/104) = 3	2(42/44) = 2
Someplace else	164(246/410) = 98	89(77/166) = 41	46(50/96) = 24	19(85/104) = 16	10(34/44) = 8
Total	$\Sigma = 281$	$\Sigma = 101$	$\Sigma = 63$	$\Sigma = 44$	$\Sigma = 20$

44 + 20) prediction errors using information about the victim-offender relationship to predict location of assault.

Goodman and Kruskal's tau is a measure of the reduction in prediction errors achieved by using knowledge of the independent variable—which, again, in our example is the victim-offender relationship. Equation 13.3 presents the general formula for calculating tau.

$$\tau = \frac{\begin{pmatrix}\text{number of errors}\\ \text{without knowledge of}\\ \text{independent variable}\end{pmatrix} - \begin{pmatrix}\text{number of errors}\\ \text{with knowledge of}\\ \text{independent variable}\end{pmatrix}}{\text{number of errors without knowledge of independent variable}}$$

Equation 13.3

For our example, tau is equal to 0.1886. If we multiply this proportion by 100%, we can discern that knowledge of the victim-offender relationship reduced our prediction errors by 18.86%, which implies a weak to moderate relationship between victim-offender relationship and location of assault.

Working It Out

$$\tau = \frac{281 - 228}{281} = 0.1886$$

Lambda (λ) is a measure of association that is conceptually very similar to Goodman and Kruskal's tau in that it is a PRE measure. However, rather than using the proportional distribution of cases to determine prediction errors, lambda uses the mode of the dependent variable. We begin with the naive decision rule, placing all possible observations in the modal category of the dependent variable and counting as errors the number of cases that do not belong in that modal category. We then use information about the value of the independent variable (the informed decision rule), making assignments of cases based on the mode of the dependent variable within each category of the independent variable.

Equation 13.4 shows that lambda is calculated in a manner similar to that used to calculate tau.

$$\lambda = \frac{\begin{pmatrix}\text{number of errors}\\ \text{using mode of}\\ \text{dependent variable}\end{pmatrix} - \begin{pmatrix}\text{number of errors}\\ \text{using mode of}\\ \text{dependent variable}\\ \text{by level of}\\ \text{independent variable}\end{pmatrix}}{\text{number of errors using mode of dependent variable}}$$

Equation 13.4

The calculation of lambda is less tedious, since we use only information on the modal category overall and then within each level of the independent variable. Without knowledge of the victim-offender relationship, we would assign all 410 cases to the assaulted-someplace-else category, resulting in $410 - 164 = 246$ classification errors.

What about the number of classification errors when we use knowledge of the victim-offender relationship? For assaults where the offender was a stranger, we would assign all 166 cases to the assaulted-someplace-else category, resulting in $166 - 89 = 77$ classification errors. For assaults where the offender was an acquaintance or friend, we would assign all 96 cases to the assaulted-someplace-else category, resulting in $96 - 46 = 50$ classification errors. All 104 partner offenders and 44 relative offenders would both be assigned to the home category, resulting in $104 - 77 = 27$ and $44 - 31 = 13$ classification errors, respectively. We have a sum of 167 prediction errors when we use knowledge of the victim-offender relationship, compared to 246 prediction errors made without any knowledge of the victim-offender relationship. The value of lambda is 0.3211, meaning that knowledge of the modal location of assault for each type of victim-offender relationship reduces our errors in predicting location of assault by 32.11%.

Working It Out

$$\lambda = \frac{246 - 167}{246} = 0.3211$$

As can be seen from our example, different measures of association may lead to somewhat different interpretations of the relationship between two variables. This occurs because different measures use different strategies in coming to a conclusion about that relationship. Which is the best measure of association for assessing the strength of the relationship between two nominal-level variables? Researchers often prefer the two PRE measures—tau and lambda—over phi and *V,* since PRE measures have direct interpretations of values that fall between 0 and 1. However, to use PRE measures, a researcher must assume that one measure (the independent variable) affects a second (the dependent variable). Of tau and lambda, tau is often defined as the better measure of association for two reasons. First, if the modal category of the dependent variable is the same for all categories of the independent variable, lambda will have a value of 0, implying that there is no relationship between the two variables. Since tau relies on the marginal distributions of observations both overall and within each category of

the independent variable, tau can still detect a relationship between the independent and the dependent variables. Second, and this is again related to the marginal distributions, the value of lambda is sensitive to marginal totals (i.e., row or column totals). When row or column totals are not approximately equal, the value of lambda may be artificially high or low. The reliance on marginal distributions in the calculation of tau allows that measure of association to account for the size of the marginal totals directly and causes it not to be as sensitive to differences in marginal totals.

Statistical Significance of Measures of Association for Nominal Variables

The statistical significance of each of the nominal measures of association just discussed can be assessed with the results of a chi-square test. When the chi-square statistic has a value of 0, each of the four coefficients will also have a value of 0. The null hypothesis for each of the four coefficients is simply that the coefficient is equal to 0. The research hypothesis is simply that the coefficient is not equal to 0.

We illustrate the steps of a hypothesis test for tau and lambda, using the data on victim-offender relationship and location of assault.

Assumptions:

Level of Measurement: Nominal scale.

Population Distribution: No assumption made.

Sampling Method: Independent random sampling.

Sampling Frame: Adults aged 18 years and older in the state of Illinois.

Hypotheses:

H_0: There is no association between victim-offender relationship and location of assault ($\tau_p = 0$).

H_1: There is an association between victim-offender relationship and location of assault ($\tau_p \neq 0$).

or

H_0: There is no association between victim-offender relationship and location of assault ($\lambda_p = 0$).

H_1: There is an association between victim-offender relationship and location of assault ($\lambda_p \neq 0$).

The Sampling Distribution Since we are testing for a relationship between two nominal-level variables, we use the chi-square distribution, where degrees of freedom $= (r - 1)(c - 1) = (4 - 1)(4 - 1) = 9$.

Table 13.9 Observed Frequencies and Expected Frequencies for Victim-Offender Relationship and Location of Assault

VICTIM-OFFENDER RELATIONSHIP	LOCATION OF ASSAULT				
	Home	Neighborhood	Work	Someplace Else	Row total
Stranger	$f_o = 10$ $f_e = 56.2780$	$f_o = 49$ $f_e = 31.1756$	$f_o = 18$ $f_e = 12.1463$	$f_o = 89$ $f_e = 66.4000$	166
Acquaintance/ friend	$f_o = 21$ $f_e = 32.5463$	$f_o = 22$ $f_e = 18.0293$	$f_o = 7$ $f_e = 7.0244$	$f_o = 46$ $f_e = 38.4000$	96
Partner	$f_o = 77$ $f_e = 35.2585$	$f_o = 5$ $f_e = 19.5317$	$f_o = 3$ $f_e = 7.6098$	$f_o = 19$ $f_e = 41.6000$	104
Relative	$f_o = 31$ $f_e = 14.9171$	$f_o = 1$ $f_e = 8.2634$	$f_o = 2$ $f_e = 3.2195$	$f_o = 10$ $f_e = 17.6000$	44
Column total	139	77	30	164	410

Table 13.10 Calculations of Chi-Square for Victim-Offender Relationship and Location of Assault

VICTIM-OFFENDER RELATIONSHIP	LOCATION OF ASSAULT	f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
Stranger	Home	10	56.2780	−46.2780	2141.6578	38.0549
Stranger	Neighborhood	49	31.1756	17.8244	317.7089	10.1909
Stranger	Work	18	12.1463	5.8537	34.2653	2.8210
Stranger	Someplace else	89	66.4000	22.6000	510.7600	7.6922
Friend	Home	21	32.5463	−11.5463	133.3180	4.0963
Friend	Neighborhood	22	18.0293	3.9707	15.7667	0.8745
Friend	Work	7	7.0244	−0.0244	0.0006	0.0001
Friend	Someplace else	46	38.4000	7.6000	57.7600	1.5042
Partner	Home	77	35.2585	41.7415	1742.3498	49.4164
Partner	Neighborhood	5	19.5317	−14.5317	211.1705	10.8117
Partner	Work	3	7.6098	−4.6098	21.2499	2.7924
Partner	Someplace else	19	41.6000	−22.6000	510.7600	12.2779
Other relative	Home	31	14.9171	16.0829	258.6605	17.3399
Other relative	Neighborhood	1	8.2634	−7.2634	52.7572	6.3844
Other relative	Work	2	3.2195	−1.2195	1.4872	0.4619
Other relative	Someplace else	10	17.6000	−7.6000	57.7600	3.2818
						$\Sigma = 168.0005$

Significance Level and Rejection Region We use the conventional 5% significance level for this example. From Appendix 2, we see that the critical value of chi-square associated with a significance level of 5% and df = 9 is 16.919. If the calculated chi-square statistic is greater than 16.919, we will reject the null hypotheses and conclude that the association between victim-offender relationship and location of assault is statistically significant.

The Test Statistic The chi-square statistic for the data in Table 13.7 is 168.001. See Table 13.9 for the expected and observed frequencies and Table 13.10 for the detailed calculations.

The Decision Since our calculated chi-square statistic of 168.001 is much larger than our critical chi-square of 16.919, we reject the null hypotheses and conclude that there is a statistically significant relationship between victim-offender relationship and location of assault.

Measures of Association for Ordinal-Level Variables

The preceding discussion described several measures of association for nominal variables, where there is no rank ordering of the categories of each variable. With ordinal-level variables, we can use the ordering of the categories to measure whether there is a positive or a negative relationship between two variables. A positive relationship would be indicated by higher ranks on one variable corresponding to higher ranks on a second variable. A negative relationship would be indicated by higher ranks on one variable corresponding to lower ranks on a second variable. The measures of association for ordinal-level variables all have values that range from −1.0 to +1.0. A value of −1.0 indicates a perfect negative relationship, a value of +1.0 indicates a perfect positive relationship, and a value of 0.0 indicates no relationship between the two variables.

Table 13.11 illustrates these variations in the strength of the relationship between two ordinal variables with a hypothetical distribution of 450 cases. Part a presents a pattern of no association between the two variables. Since the cases are evenly distributed across all the cells of the table, knowledge of the level of one ordinal variable does not provide any information about the level of the second ordinal variable. Parts b and c show perfect negative and positive relationships, respectively, where knowledge of the level of one ordinal variable determines, without error, the level of the second ordinal variable.

The calculation of ordinal measures of association is tedious to perform by hand. When doing data analysis, you would likely rely on a statistical

Table 13.11 Hypothetical Distribution of 450 Cases for Two Ordinal Variables

(a) Measure of Association = 0.0

	VARIABLE 2			
VARIABLE 1	**Low**	**Medium**	**High**	**Row total**
Low	50	50	50	150
Medium	50	50	50	150
High	50	50	50	150
Column total	150	150	150	450

(b) Measure of Association = −1.0

	VARIABLE 2			
VARIABLE 1	**Low**	**Medium**	**High**	**Row total**
Low	0	0	150	150
Medium	0	150	0	150
High	150	0	0	150
Column total	150	150	150	450

(c) Measure of Association = +1.0

	VARIABLE 2			
VARIABLE 1	**Low**	**Medium**	**High**	**Row total**
Low	150	0	0	150
Medium	0	150	0	150
High	0	0	150	150
Column total	150	150	150	450

software package to perform the calculations for you. Most common statistical software packages will compute the measures of association for ordinal variables described here (see, for example, the computer exercises at the end of this chapter). The following discussion is intended to help you understand how these various measures are calculated.

There are four common measures of association for ordinal variables: **gamma (γ), Kendall's τ_b, Kendall's τ_c,** and **Somers' *d*.** Common to all four is the use of **concordant pairs** and **discordant pairs of observations.** The logic behind using concordant and discordant pairs of observations is that we take each possible pair of observations in a data set and compare the relative ranks of the two observations on the two variables examined. Concordant pairs are those pairs of observations for which the rankings are consistent: One observation is ranked high on both variables, while the other observation is ranked low on both variables. For example, one observation is ranked 1 (of five ranked categories)

on the first variable and 2 (of five ranked categories) on the second variable, while the other observation is ranked 4 on the first variable and 3 on the second variable. Discordant pairs refer to those pairs of observations for which the rankings are inconsistent: One observation is ranked high on the first variable and low on the second variable, while the other observation is ranked low on the first variable and high on the second variable. For example, one observation is ranked 1 on the first variable and 5 on the second variable, while the other observation is ranked 4 on the first variable and 2 on the second variable. A pair of observations that has the same rank on one or both of the variables is called a **tied pair of observations (tie).**[4] Somers' d is the only one of the four measures of association for ordinal variables for which specification of the dependent and the independent variables is required. The value of d depending on which variable is specified as the dependent variable. To simplify the following discussion, the examples we present in the next section define one variable as the dependent and the other as the independent variable.

How do we decide whether a pair of observations is a concordant pair, a discordant pair, or a tied pair? Let's look at the determination of concordant, discordant, and tied pairs for the data presented in Table 13.12, which represents a cross-tabulation of two ordinal variables, each with three categories: low, medium, and high.

Table 13.12 Cross-Tabulation of Two Ordinal Variables

	DEPENDENT VARIABLE		
INDEPENDENT VARIABLE	**Low**	**Medium**	**High**
Low	Cell A 12	Cell B 4	Cell C 3
Medium	Cell D 5	Cell E 10	Cell F 6
High	Cell G 3	Cell H 5	Cell I 14

[4]All the measures of association for ordinal variables that we discuss here are for grouped data that can be represented in the form of a table. In Chapter 14, we discuss another measure of association for ordinal variables—Spearman's r (r_s)—that is most useful in working with ungrouped data, such as information on individuals. The difficulty we confront when using Spearman's r on grouped data is that the large number of tied pairs of observations complicates the calculation of this measure of association. Spearman's r is a more appropriate measure of association when we have ordinal variables with a large number of ranked categories for individual cases or when we take an interval-level variable and rank order the observations (see Chapter 14).

We begin by determining the concordant pairs—those pairs of observations that have consistent relative rankings. Let's start with Cell A. We remove from consideration the row and column that Cell A is located in, since the cases in the same row or column will have the same ranking on the independent and dependent variables, respectively, and thus represent ties. We then look for cases located *below* and to the *right* of the cell of interest. For Cell A, the cells we will use to determine concordant pairs are Cells E, F, H, and I, since the ranks are consistently lower on both the independent and the dependent variables. To determine the number of pairs of observations that are concordant for observations in Cell A, we begin by summing the number of observations in Cells E, F, G, and I: 10 + 6 + 5 + 14 = 35. This tells us that for a single observation in Cell A, there are 35 concordant pairs of observations. Since there are 12 observations in Cell A, we multiply the number of cases in Cell A (12) by the sum of the cases in Cells E, F, H, and I. For Cell A, there are 420 concordant pairs.

Working It Out

$$12(10 + 6 + 5 + 14) = 420$$

Continuing to work across the first row of Table 13.12, we move to Cell B. The cells located below and to the right of Cell B are Cells F and I, so the number of concordant pairs is 4(6 + 14) = 80. When we move to Cell C, we see there are no cells below and to the right, so we drop down to the next row and start with Cell D. The cells located below and to the right of Cell D are Cells H and I, so the number of concordant pairs is 5(5 + 14) = 95. Moving to Cell E, we see that only Cell I is below and to the right, so the number of concordant pairs is 10(14) = 140. The remaining cells in the table—F, G, H, and I—have no other cells located below and to the right, so they are not used in the calculation of concordant pairs. After calculating concordant pairs for all cells in the table, we sum these values to get the number of concordant pairs for the table. For Table 13.12, the total number of concordant pairs is 735 (that is, 420 + 80 + 95 + 140).

Working It Out

$$\begin{aligned}
\text{Cell A:} \quad & 12(10 + 6 + 5 + 14) = 420 \\
\text{Cell B:} \quad & 4(6 + 14) = 80 \\
\text{Cell D:} \quad & 5(5 + 14) = 95 \\
\text{Cell E:} \quad & 10(14) = 140
\end{aligned}$$

$$\text{Sum} = 420 + 80 + 95 + 140 = 735$$

To calculate discordant cells, we begin in the upper right corner of Table 13.12 (Cell C), locate cells that are positioned *below* and to the *left* of the cell of interest, and perform calculations similar to those for concordant pairs. Beginning with Cell C, we multiply the number of cases in Cell C by the sum of cases in Cells D, E, G, and H, which are located below and to the left of Cell C. The number of discordant pairs for Cell C is 69.

Working It Out

$3(5 + 10 + 3 + 5) = 69$

Moving from right to left in the top row of Table 13.12, we shift our attention to Cell B. The discordant pairs for Cell B are calculated by multiplying the number of cases in Cell B by the sum of cases in Cells D and G. We find the number of discordant pairs for Cell B to be $4(5 + 3) = 32$. Since there are no cells located below and to the left of Cell A, it is not used to calculate discordant pairs, and we move on to Cell F. The cells located below and to the left of Cell F are Cells G and H, so the number of discordant pairs is $6(3 + 5) = 48$. For Cell E, the only cell located below and to the left is Cell G, so the number of discordant pairs is $10(3) = 30$. There are no cells located below and to the left of Cells D, G, H, and I, so no further calculations are performed. As with the concordant pairs, we sum our discordant pairs for Table 13.12 and find the sum to be 179 (that is, $69 + 32 + 48 + 30$).

Working It Out

Cell C: $3(5 + 10 + 3 + 5) = 69$
Cell B: $4(5 + 3) = 32$
Cell F: $6(3 + 5) = 48$
Cell E: $10(3) = 30$

Sum $= 69 + 32 + 48 + 30 = 179$

To calculate ties in rank for pairs of observations, we have to consider the independent and dependent variables separately. We denote ties on the independent variable as T_X and ties on the dependent variable as T_Y. Since the independent variable is represented by the rows

in Table 13.12, the pairs of observations that will be defined as ties on the independent variable will be those cases located in the same row of Table 13.12. To calculate the number of ties in each row, we use Equation 13.5.

$$T_X = \frac{1}{2} \sum N_{\text{row}}(N_{\text{row}} - 1) \qquad \textbf{Equation 13.5}$$

where T_X is the number of ties on the independent variable and N_{row} is the row total. Equation 13.5 tells us to calculate the product of the number of observations in a row and the number of observations in a row minus 1 for all rows. We then sum the products calculated for each row and multiply the sum by ½.

For Table 13.12, the three row totals are 19 (row 1), 21 (row 2), and 22 (row 3). When we insert these values into Equation 13.5, we find the number of ties on the independent variable to be 612.

Working It Out

$$T_X = \frac{1}{2} \sum N_{\text{row}}(N_{\text{row}} - 1)$$

$$= \frac{1}{2} [(19)(19 - 1) + (21)(21 - 1) + (22)(22 - 1)]$$

$$= \frac{1}{2} [(19)(18) + (21)(20) + (22)(21)]$$

$$= \frac{1}{2} (342 + 420 + 462) = \frac{1}{2} (1{,}224)$$

$$= 612$$

The ties on the dependent variable are found in a similar manner. Since the dependent variable is represented in the columns, we perform the same type of calculation, but using column totals rather than row totals. Equation 13.6 presents the formula for calculating ties on the dependent variable.

$$T_Y = \frac{1}{2} \sum N_{\text{col}}(N_{\text{col}} - 1) \qquad \textbf{Equation 13.6}$$

In Equation 13.6, T_Y is the number of ties on the dependent variable and N_{col} is the total number of observations in the column.

In Table 13.12, the column totals are 20 (column 1), 19 (column 2), and 23 (column 3). After inserting these values into Equation 13.6, we find the number of ties on the dependent variable to be 614.

Working It Out

$$T_Y = \frac{1}{2}\sum N_{\text{col}}(N_{\text{col}} - 1)$$

$$= \frac{1}{2}[(20)(20 - 1) + (19)(19 - 1) + (23)(23 - 1)]$$

$$= \frac{1}{2}[(20)(19) + (19)(18) + (23)(22)]$$

$$= \frac{1}{2}(380 + 342 + 506) = \frac{1}{2}(1{,}228)$$

$$= 614$$

Gamma

Once we have calculated the numbers of concordant pairs and discordant pairs, gamma (γ) is the simplest of the ordinal measures of association to calculate, since it does not use information about ties in rank. Gamma has possible values that range from -1.0 to $+1.0$. Gamma may also be interpreted as a PRE measure: We can interpret the value of gamma as indicating the proportional reduction in errors in predicting the dependent variable, based on information about the independent variable.

Equation 13.7 presents the formula for calculating gamma. Gamma is the difference between the number of concordant (C) and discordant (D) pairs, ($C - D$), divided by the sum of the concordant and discordant pairs, ($C + D$).

$$\gamma = \frac{C - D}{C + D} \qquad \textbf{Equation 13.7}$$

For the data in Table 13.12, gamma is equal to 0.6083. The positive value of gamma tells us that as we move from lower ranked to higher ranked categories on the independent variable, the category of the dependent variable also tends to increase. In regard to the relative strength of the relationship, a value of 0.6083 suggests a strong relationship between the independent and dependent variables, since knowledge of the independent variable reduces our errors in predicting the dependent variable by 60.83%.

Working It Out

$$\gamma = \frac{C - D}{C + D}$$
$$= \frac{735 - 179}{735 + 179}$$
$$= \frac{556}{914}$$
$$= 0.6083$$

Kendall's τ_b and τ_c

Kendall's tau measures—τ_b and τ_c—also assess the strength of association between two ordinal variables.[5] The two measures are conceptually very similar in that they use information about concordant and discordant pairs of observations. But they also utilize information about tied pairs on both the independent and the dependent variables. Both tau measures have possible values ranging from −1.0 to +1.0. There are two important differences between τ_b and τ_c: First, τ_b should be applied *only* to a table where the number of rows is equal to the number of columns; τ_c should be applied to a table where the number of rows is not equal to the number of columns. When the number of rows is equal to the number of columns, τ_c will have a value close to that of τ_b. Second, τ_b may be interpreted as a PRE measure, but τ_c may not. The differences in the application and interpretation of each measure suggest that knowing the dimensions of the table is important in deciding which measure is most appropriate.

Equations 13.8 and 13.9 present the formulas for calculating τ_b and τ_c, respectively.

$$\tau_b = \frac{C - D}{\sqrt{[N(N-1)/2 - T_X][N(N-1)/2 - T_Y]}} \qquad \textbf{Equation 13.8}$$

In Equation 13.8, C and D represent the concordant and the discordant pairs, respectively; N represents the total number of cases; T_X represents the number of ties on the independent variable; and T_Y represents the number of ties on the dependent variable.

[5]These two tau measures are different from Goodman and Kruskal's tau, which measures the strength of association between two nominal variables.

Let's return to the data presented in Table 13.12. We have already calculated the number of concordant pairs to be 735, the number of discordant pairs to be 179, the total number of cases to be 62, the number of ties on the independent variable to be 612, and the number of ties on the dependent variable to be 614. After inserting these values into Equation 13.6, we find τ_b to be 0.4351. This indicates that knowledge of the independent variable reduces our prediction errors by 43.51%.

Working It Out

$$\tau_b = \frac{C - D}{\sqrt{[N(N-1)/2 - T_X][N(N-1)/2 - T_Y]}}$$

$$= \frac{735 - 179}{\sqrt{[62(62-1)/2 - 612][62(62-1)/2 - 614]}}$$

$$= \frac{556}{\sqrt{[1{,}891 - 612][1{,}891 - 614]}}$$

$$= \frac{556}{\sqrt{(1{,}279)(1{,}277)}}$$

$$= 0.4351$$

Equation 13.9 presents the formula for calculating τ_c. We do not calculate τ_c for Table 13.12, since the number of rows is equal to the number of columns. We do, however, illustrate its calculation below with another example.

$$\tau_c = \frac{C - D}{\frac{1}{2}N^2[(m-1)/m]} \qquad \textbf{Equation 13.9}$$

where $m = \min(r, c)$

In Equation 13.9, C and D represent the concordant and the discordant pairs, respectively; N represents the total number of cases; and m is the smaller of the number of rows (r) and the number of columns (c). Suppose, for example, that we had a table with five rows ($r = 5$) and four columns ($c = 4$). The number of columns is smaller than the number of rows, so m would be 4.

Somers' *d*

The fourth measure of association for ordinal variables that we present here—Somers' *d*—is similar to the tau measures, but instead of using information about ties on both the independent and the dependent variables, Somers' *d* uses information on ties on only the independent variable. It is important to remember that the statistic you get for Somers' *d* may vary, depending on which variable is defined as the dependent variable. The formula for calculating Somers' *d* is given in Equation 13.10.

$$d_{YX} = \frac{C - D}{N(N - 1)/2 - T_X} \qquad \textbf{Equation 13.10}$$

In Equation 13.10, where *C, D, N,* and T_X represent the concordant pairs, the discordant pairs, the total number of cases, and the number of ties on the independent variable, respectively. The subscript *YX* on *d* denotes the dependent and the independent variables, in order.

For Table 13.12, we have already calculated values for *C, D, N,* and T_X. After inserting these values into Equation 13.10, we find Somers' *d* to be 0.4347.

Working It Out

$$\begin{aligned} d_{YX} &= \frac{C - D}{N(N - 1)/2 - T_X} \\ &= \frac{735 - 179}{62(62 - 1)/2 - 612} \\ &= \frac{556}{1{,}891 - 612} \\ &= \frac{556}{1{,}279} \\ &= 0.4347 \end{aligned}$$

A Substantive Example: Affectional Identification with Father and Level of Delinquency

Table 9.14 presented a cross-tabulation of two ordinal variables: affectional identification with father and delinquency. Identification with father was determined by the youth's responses to a question about how much they wanted to grow up and be like their fathers. The responses were classified into five ordered categories: in every way, in most

Table 13.13 Affectional Identification with Father by Number of Delinquent Acts

AFFECTIONAL IDENTIFICATION WITH FATHER	NUMBER OF DELINQUENT ACTS			
	None	One	Two or more	Row total
In every way	Cell A 77	Cell B 25	Cell C 19	121
In most ways	Cell D 263	Cell E 97	Cell F 44	404
In some ways	Cell G 224	Cell H 97	Cell I 66	387
In just a few ways	Cell J 82	Cell K 52	Cell L 38	172
Not at all	Cell M 56	Cell N 30	Cell O 52	138
Column total	702	301	219	1,222

ways, in some ways, in just a few ways, and not at all. Delinquent acts were classified into three ordered categories: none, one, and two or more. The data came from the Richmond Youth Survey report, and the distribution of cases presented refers only to the white males who responded to the survey.[6] We reproduce this cross-tabulation in Table 13.13.

In our earlier analysis of the data in this table (see Chapter 9), we found a statistically significant relationship between affectional identification with father and delinquency. However, the chi-square statistic told us nothing about the direction of the effect or the strength of the relationship between these two variables. We can use the measures of association for ordinal variables to test the strength of the relationship between identification with father and level of delinquency.

We begin by calculating the numbers of concordant pairs, discordant pairs, and tied pairs of observations. The number of concordant pairs of observations is 201,575; the number of discordant pairs is 125,748; the number of pairs tied on the independent variable is 187,516; and the number of pairs tied on the dependent variable is 315,072.

[6]David F. Greenberg, "The Weak Strength of Social Control Theory," *Crime and Delinquency* 45:1 (1999): 66–81.

Working It Out

Concordant Pairs:

Cell A: $77(97 + 44 + 97 + 66 + 52 + 38 + 30 + 52) = 36{,}652$

Cell B: $25(44 + 66 + 38 + 52) = 5{,}000$

Cell D: $263(97 + 66 + 52 + 38 + 30 + 52) = 88{,}105$

Cell E: $97(66 + 38 + 52) = 15{,}132$

Cell G: $224(52 + 38 + 30 + 52) = 38{,}528$

Cell H: $97(38 + 52) = 8{,}730$

Cell J: $82(30 + 52) = 6{,}724$

Cell K: $52(52) = 2{,}704$

$$\text{Sum} = 36{,}652 + 5{,}000 + 88{,}105 + 15{,}132 + 38{,}528 + 8{,}730 + 6{,}724 + 2{,}704$$

$$= 201{,}575$$

Discordant Pairs:

Cell C: $19(263 + 97 + 224 + 97 + 82 + 52 + 56 + 30) = 17{,}119$

Cell B: $25(263 + 224 + 82 + 56) = 15{,}625$

Cell F: $44(224 + 97 + 82 + 52 + 56 + 30) = 23{,}804$

Cell E: $97(224 + 82 + 56) = 35{,}114$

Cell I: $66(82 + 52 + 56 + 30) = 14{,}520$

Cell H: $97(82 + 56) = 13{,}386$

Cell L: $38(56 + 30) = 3{,}268$

Cell K: $52(56) = 2{,}912$

$$\text{Sum} = 17{,}119 + 15{,}625 + 23{,}804 + 35{,}114 + 14{,}520 + 13{,}386 + 3{,}268 + 2{,}912$$

$$= 125{,}748$$

Pairs Tied on the Independent Variable:

$$T_X = (\tfrac{1}{2})[(121)(120) + (404)(403) + (387)(386) + (172)(171) + (138)(137)]$$
$$= (\tfrac{1}{2})(14{,}520 + 162{,}812 + 149{,}382 + 29{,}412 + 18{,}906)$$
$$= (\tfrac{1}{2})(375{,}032)$$
$$= 187{,}516$$

Pairs Tied on the Dependent Variable:

$$T_Y = (\tfrac{1}{2})[(702)(701) + (301)(300) + (219)(218)]$$
$$= (\tfrac{1}{2})(492{,}102 + 90{,}300 + 47{,}742)$$
$$= (\tfrac{1}{2})(630{,}144)$$
$$= 315{,}072$$

After calculating the concordant pairs, discordant pairs, and pairs tied on the independent and dependent variables, we can calculate the measures of association for ordinal variables. We find the value of gamma to be 0.2317. Don't be confused by the fact that for affectional identification movement from lower to higher ordered categories represents movement from more to less identification with the father. Substantively, what this value of gamma tells us is that as the level of affectional identification with father decreases (i.e., as we move down the rows of the table), the youth are likely to report higher levels of delinquency. The value of gamma also indicates that we reduce our prediction errors about level of delinquency by 23.17% when we use information about the level of affectional identification with father. If affectional identification in this example had been measured from less to more identification with father (rather than more to less identification), gamma would have been negative. As a general rule, it is important to look carefully at the ordering of the categories of your measure in order to make a substantive interpretation of your result.

Working It Out

$$\gamma = \frac{C - D}{C + D}$$

$$= \frac{201{,}575 - 125{,}748}{201{,}575 + 125{,}748}$$

$$= \frac{75{,}827}{327{,}323}$$

$$= 0.2317$$

Recall that there are two tau measures: τ_b and τ_c. If the number of rows were equal to the number of columns, then we would use τ_b. Since the number of rows is different from the number of columns in Table 13.13, we use τ_c. For the data presented in Table 13.13, τ_c has a value of 0.1523, meaning that as the level of affectional identification with father decreases, the level of delinquency increases. However, since τ_c is not a PRE measure, we cannot interpret this result in terms of proportional reduction in error.

Working It Out

$$\tau_c = \frac{C - D}{\frac{1}{2}N^2\left(\frac{m - 1}{m}\right)}, \quad \text{where } m = \min(r, c) = \min(5, 3) = 3$$

$$= \frac{201{,}575 - 125{,}748}{\frac{1}{2}(1{,}222)^2\left(\frac{3 - 1}{3}\right)}$$

$$= \frac{75{,}827}{497{,}761.3333}$$

$$= 0.1523$$

Our third measure of association for ordinal variables, Somers' *d*, has a value of 0.1358. The interpretation is the same as that for gamma and τ_c: Lower levels of affectional identification with father are associated with higher levels of delinquency. In this case, knowledge of level of affectional identification with father reduces our prediction errors about level of delinquency by 13.58%.

Working It Out

$$d_{YX} = \frac{C - D}{N(N-1)/2 - T_X}$$

$$= \frac{201{,}575 - 125{,}748}{[(1{,}222)(1{,}222 - 1)/2] - 187{,}516}$$

$$= \frac{75{,}827}{558{,}515}$$

$$= 0.1358$$

Note on the Use of Measures of Association for Ordinal Variables

As illustrated in our example, the values for gamma, Kendall's tau measures, and Somers' *d* will generally not be the same. The difference in values can be attributed primarily to whether the measure accounts for tied pairs of observations. Gamma does not account for tied pairs of observations and thus is sometimes criticized for overestimating the strength of association between two ordinal variables. Somers' *d* accounts for only the pairs of observations tied on the independent variable, while Kendall's tau measures account for tied pairs of observations on both variables.

Which of these measures is best to use in which situations? As in our discussion of measures of association for nominal variables, to begin to address this question, we need to consider the dimensions of the table and our desire for a PRE measure. If the number of rows is equal to the number of columns, then τ_b is likely the best overall measure of association for two reasons: First, it has a PRE interpretation, meaning that values falling between 0 and 1 have direct interpretations in terms of reduction of error. Second, since τ_b accounts for pairs of observations tied on both the independent and the dependent variables, it will provide a more conservative estimate than gamma. If the number of rows is not equal to the number of columns, Somers' *d* is sometimes considered a better measure of association than τ_c, since it has a PRE interpretation and τ_c does not. Somers' *d* offers the additional advantage of being an appropriate measure of association for those situations where we have clearly defined independent and dependent variables.

Statistical Significance of Measures of Association for Ordinal Variables

Each of the four measures of association for ordinal variables can be tested for statistical significance with a *z*-test. The general formula for calculating the *z*-score is given in Equation 13.11, where we divide

the measure of association by the standard error of the measure of association.

$$z = \frac{\text{measure of association}}{\text{standard error of measure of association}}$$

Equation 13.11

What will differ for each of the measures of association for ordinal variables is the calculation of the standard error. Equations 13.12, 13.13, and 13.14 present *approximate* standard errors for gamma, Kendall's tau measures, and Somers' *d,* respectively.[7]

$$\hat{\sigma}_\gamma = \sqrt{\frac{4(r+1)(c+1)}{9N(r-1)(c-1)}}$$

Equation 13.12
Approximate Standard Error for Gamma

$$\hat{\sigma}_\tau = \sqrt{\frac{4(r+1)(c+1)}{9Nrc}}$$

Equation 13.13
Approximate Standard Error for Kendall's Tau Measures

$$\hat{\sigma}_d = \sqrt{\frac{4(r^2-1)(c+1)}{9Nr^2(c-1)}}$$

Equation 13.14
Approximate Standard Error for Somers' *d*

In all three equations, N is the total number of observations, r is the number of rows, and c is the number of columns in the table.

Assumptions:

Level of Measurement: Ordinal scale.

Population Distribution: Normal distribution for the relationship examined (relaxed because N is large).

Sampling Method: Independent random sampling.

Sampling Frame: High school–age white males in Richmond, California, in 1965.

Hypotheses:

H_0: There is no association between affectional identification with father and delinquency ($\gamma_p = 0$).

H_1: There is an association between affectional identification with father and delinquency ($\gamma_p \neq 0$).

---

[7]For a more detailed discussion of these issues, see Jean Dickson Gibbons, *Nonparametric Measures of Association* (Newbury Park, CA: Sage, 1993).

or

H_0: There is no association between affectional identification with father and delinquency ($\tau_{c(p)} = 0$).

H_1: There is an association between affectional identification with father and delinquency ($\tau_{c(p)} \neq 0$).

or

H_0: There is no association between affectional identification with father and delinquency ($d_p = 0$).

H_1: There is an association between affectional identification with father and delinquency ($d_p \neq 0$).

The Sampling Distribution We use the normal distribution to test whether the measures of ordinal association differ significantly from 0. As with our earlier examples using a normal sampling distribution, the N of cases must be large in order for us to relax the normality assumption. When examining the relationship between two ordinal-level variables, we recommend a sample of at least 60 cases.

Significance Level and Rejection Region We use the conventional 5% significance level for our example. From Appendix 3, we can determine that the critical values for z are ± 1.96. If the calculated z-score is greater than 1.96 or less than -1.96, we will reject the null hypotheses and conclude that the measure of association between affectional identification with father and delinquency is significantly different from 0.

The Test Statistic Since we have three different measures of association—γ, τ_c, and d—we need to calculate three separate test statistics. We first need to calculate the approximate standard error for gamma, using Equation 13.12. We find the standard error for gamma to be 0.0330.

Working It Out

$$\hat{\sigma}_\gamma = \sqrt{\frac{4(r+1)(c+1)}{9N(r-1)(c-1)}}$$

$$= \sqrt{\frac{4(5+1)(3+1)}{(9)(1{,}222)(5-1)(3-1)}}$$

$$= \sqrt{\frac{96}{87{,}984}}$$

$$= \sqrt{0.00109}$$

$$= 0.0330$$

Using the standard error for gamma, we then calculate the *z*-score using Equation 13.11. In our example, we find the *z*-score for gamma to be 7.0212.

Working It Out

$$z = \frac{\gamma}{\hat{\sigma}_\gamma}$$
$$= \frac{0.2317}{0.0330}$$
$$= 7.0212$$

Turning to τ_c, we calculate the standard error using Equation 13.13. For our example, the standard error for τ_c is 0.0241.

Working It Out

$$\hat{\sigma}_\tau = \sqrt{\frac{4(r+1)(c+1)}{9Nrc}}$$
$$= \sqrt{\frac{4(5+1)(3+1)}{(9)(1{,}222)(5)(3)}}$$
$$= \sqrt{\frac{96}{164{,}970}}$$
$$= \sqrt{0.00058}$$
$$= 0.0241$$

Using the standard error for τ_c and Equation 13.11, we find the *z*-score to be 6.3195.

Working It Out

$$z = \frac{\tau_c}{\hat{\sigma}_\tau}$$
$$= \frac{0.1523}{0.0241}$$
$$= 6.3195$$

For Somers' *d,* we follow the same process, calculating the standard error for *d* and then using the standard error to calculate the *z*-score for *d*. For our example, the standard error for *d* is 0.0264 and the corresponding *z*-score is 5.1439.

Working It Out

$$\hat{\sigma}_d = \sqrt{\frac{4(r^2 - 1)(c + 1)}{9Nr^2(c - 1)}}$$

$$= \sqrt{\frac{4(5^2 - 1)(3 + 1)}{(9)(1,222)(5^2)(3 - 1)}}$$

$$= \sqrt{\frac{384}{549,900}}$$

$$= \sqrt{0.00070}$$

$$= 0.0264$$

Working It Out

$$z = \frac{d}{\hat{\sigma}_d}$$

$$= \frac{0.1358}{0.0264}$$

$$= 5.1439$$

The Decision All three of the calculated *z*-scores are greater than 1.96, meaning that we reject the null hypotheses and conclude in the case of each test that there is a statistically significant relationship between affectional identification with father and delinquency.

Choosing the Best Measure of Association for Nominal- and Ordinal-Level Variables

Because we have covered so many different measures in this chapter, we thought it would be useful to recap them in a simple table that can be used in deciding which measure of association is appropriate

Table 13.14 Summary of Measures of Association for Nominal and Ordinal Variables

MEASURE OF ASSOCIATION	LEVEL OF MEASUREMENT	PRE MEASURE?	DIMENSIONS OF TABLE (ROWS BY COLUMNS)
ϕ	Nominal	No	2×2
V	Nominal	No	Any size
λ	Nominal	Yes	Any size
τ	Nominal	Yes	Any size
γ	Ordinal	Yes	Any size
τ_b	Ordinal	Yes	Number of rows = Number of columns
τ_c	Ordinal	No	Number of rows ≠ Number of columns
d	Ordinal	Yes	Any size

for which specific research problem. Table 13.14 presents summary information on the measures of association for nominal and ordinal variables discussed in this chapter. The first column of Table 13.14 gives the measure of association, the second column notes the appropriate level of measurement for the two variables, the third column tells whether the measure of association is also a PRE measure, and the fourth column lists any restrictions on the size of the table used in the analysis. Thus, for any given pair of nominal or ordinal variables, you should be able to determine which measure of association best suits your needs.

Chapter Summary

Measures of association for nominal and ordinal variables allow researchers to go beyond a simple chi-square test for independence between two variables and assess the strength of the relationship. The measures of association discussed in this chapter are the most commonly used measures of association for nominal and ordinal variables.

Two of the measures of association for nominal variables are based on the value of the chi-square statistic. **Phi (ϕ)** adjusts the value of chi-square by taking into account the size of the sample, but is useful only for 2×2 tables. **Cramer's *V*** is also based on the value of the chi-square statistic, but makes an additional adjustment for the numbers of rows and columns in the table. One of the difficulties with the interpretation

of phi and V is that a value that falls between 0 and 1 does not have a precise interpretation. We can infer that as values approach 0, there is a weak relationship between the two variables. Similarly, as values approach 1, there is a strong (or near perfect) relationship between the two variables.

Goodman and Kruskal's tau and **lambda** are measures of association that are not based on the value of the chi-square statistic and instead use different decision rules for classifying cases. Tau relies on the proportional distribution of cases in a table, while lambda relies on the modal values of the dependent variable overall and within each level or category of the independent variable. Tau and lambda offer an improvement over phi and V in that a value between 0 and 1 can be interpreted directly as the proportional reduction in errors made by using information about the independent variable. More generally, this characteristic is called **proportional reduction in error,** or **PRE.** PRE measures tell us how much knowledge of one measure helps to reduce the errors we make in defining the values of a second measure. Both measures require that we define at the outset which variable is the **dependent variable** and which variable is the **independent variable.** The dependent variable is the outcome variable—the phenomenon that we are interested in explaining. As it is dependent on other variables, it is influenced—or we expect it to be influenced—by other variables. The variables that affect, or influence, the dependent variable are referred to as the independent variables.

There are four common measures of association for ordinal variables: **gamma (γ), Kendall's τ_b and τ_c**, and **Somers' *d*.** Measures of association for ordinal variables are all based on **concordant pairs** and **discordant pairs** of observations. Concordant pairs are pairs of observations that have consistent rankings on the two variables (e.g., high on both variables or low on both variables), while discordant pairs are those pairs of observations that have inconsistent rankings on the two variables (e.g., high on one variable and low on the other variable). Gamma uses information only on the concordant and discordant pairs of observations. The remaining measures of association—Kendall's tau measures and Somers' d—use information about pairs of observations that have tied rankings. All four of the measures of association for ordinal variables discussed in this chapter have values ranging from -1.0 to 1.0, where a value of -1.0 indicates a perfect negative relationship (i.e., as we increase the value of one variable, the other variable decreases), a value of 1.0 indicates a perfect positive relationship (i.e., as we increase the value of one variable, the other variable also increases), and a value of 0.0 indicates no relationship between the two variables. Gamma (γ), Kendall's τ_b, and Somers' d all have PRE interpretations.

Key Terms

concordant pairs of observations Pairs of observations that have consistent rankings on two ordinal variables.

Cramer's *V* A measure of association for two nominal variables that adjusts the chi-square statistic by the sample size. *V* is appropriate when at least one of the nominal variables has more than two categories.

dependent variable The outcome variable; the phenomenon that we are interested in explaining. It is dependent on other variables in the sense that it is influenced—or we expect it to be influenced—by other variables.

discordant pairs of observations Pairs of observations that have inconsistent rankings on two ordinal variables.

gamma (γ) PRE measure of association for two ordinal variables that uses information about concordant and discordant pairs of observations within a table. Gamma has a standardized scale ranging from −1.0 to 1.0.

Goodman and Kruskal's tau (τ) PRE measure of association for two nominal variables that uses information about the proportional distribution of cases within a table. Tau has a standardized scale ranging from 0 to 1.0. For this measure, the researcher must define the independent and dependent variables.

independent variable A variable assumed by the researcher to affect or influence the dependent variable.

Kendall's τ_b PRE measure of association for two ordinal variables that uses information about concordant pairs, discordant pairs, and pairs of observations tied on both variables examined. τ_b has a standardized scale ranging from −1.0 to 1.0 and is appropriate only when the number of rows equals the number of columns in a table.

Kendall's τ_c A measure of association for two ordinal variables that uses information about concordant pairs, discordant pairs, and pairs of observations tied on both variables examined. τ_c has a standardized scale ranging from −1.0 to 1.0 and is appropriate when the number of rows is not equal to the number of columns in a table.

lambda (λ) PRE measure of association for two nominal variables that uses information about the modal category of the dependent variable for each category of the independent variable. Lambda has a standardized scale ranging from 0 to 1.0.

phi (ϕ) A measure of association for two nominal variables that adjusts the chi-square statistic by the sample size. Phi is appropriate only for nominal variables that each have two categories.

proportional reduction in error (PRE) The proportional reduction in errors made when the value of one measure is predicted using information about the second measure.

Somers' *d* PRE measure of association for two ordinal variables that uses information about concordant pairs, discordant pairs, and pairs of observations tied on the independent variable. Somers' *d* has a standardized scale ranging from −1.0 to 1.0.

tied pairs of observations (ties) Pairs of observation that have the same ranking on two ordinal variables.

Symbols and Formulas

C	Number of concordant pairs of observations
D	Number of discordant pairs of observations
N_{row}	Total number of observations for each row
N_{col}	Total number of observations for each column
T_X	Number of pairs of observations tied on the independent variable
T_Y	Number of pairs of observations tied on the dependent variable
ϕ	Phi; measure of association for nominal variables
V	Cramer's V; measure of association for nominal variables
λ	Lambda; measure of association for nominal variables
τ	Goodman and Kruskal's tau; measure of association for nominal variables
γ	gamma; measure of association for ordinal variables
τ_b	Kendall's τ_b; measure of association for ordinal variables
τ_c	Kendall's τ_c; measure of association for ordinal variables
d	Somers' d; measure of association for ordinal variables

To calculate phi (ϕ):

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

To calculate Cramer's V:

$$V = \sqrt{\frac{\chi^2}{N \times \min(r - 1,\ c - 1)}}$$

To calculate Goodman and Kruskal's tau:

$$\tau = \frac{\begin{pmatrix}\text{number of errors}\\ \text{without knowledge of}\\ \text{independent variable}\end{pmatrix} - \begin{pmatrix}\text{number of errors}\\ \text{with knowledge of}\\ \text{independent variable}\end{pmatrix}}{\text{number of errors without knowledge of independent variable}}$$

To calculate lambda:

$$\lambda = \frac{\begin{pmatrix}\text{number of errors}\\ \text{using mode of}\\ \text{dependent variable}\end{pmatrix} - \begin{pmatrix}\text{number of errors}\\ \text{using mode of}\\ \text{dependent variable}\\ \text{by level of}\\ \text{independent variable}\end{pmatrix}}{\text{number of errors using mode of dependent variable}}$$

To calculate the number of tied pairs of observations on the independent variable:

$$T_X = \frac{1}{2}\sum N_{\text{row}}(N_{\text{row}} - 1)$$

To calculate the number of tied pairs of observations on the dependent variable:

$$T_Y = \frac{1}{2}\sum N_{\text{col}}(N_{\text{col}} - 1)$$

To calculate gamma:

$$\gamma = \frac{C - D}{C + D}$$

To calculate τ_b:

$$\tau_b = \frac{C - D}{\sqrt{[N(N-1)/2 - T_X][N(N-1)/2 - T_Y]}}$$

To calculate τ_c:

$$\tau_c = \frac{C - D}{\frac{1}{2}N^2[(m-1)/m]}, \text{ where } m = \min(r, c)$$

To calculate Somers' *d:*

$$d_{YX} = \frac{C - D}{N(N-1)/2 - T_X}$$

To calculate the *z*-score:

$$z = \frac{\text{measure of association}}{\text{standard error of measure of association}}$$

To calculate the standard error for gamma:

$$\hat{\sigma}_{\gamma} = \sqrt{\frac{4(r + 1)(c + 1)}{9N(r - 1)(c - 1)}}$$

To calculate the standard error for Kendall's tau measures:

$$\hat{\sigma}_{\tau} = \sqrt{\frac{4(r + 1)(c + 1)}{9Nrc}}$$

To calculate the standard error for Somers' *d:*

$$\hat{\sigma}_{d} = \sqrt{\frac{4(r^2 - 1)(c + 1)}{9Nr^2(c - 1)}}$$

Exercises

13.1 A researcher studies the link between race of offender and death sentence decision in a state by selecting a random sample of death penalty cases over a 20-year period. The researcher finds the following distribution of death sentence decisions by race:

Race	Sentenced to Death	Not Sentenced to Death
White	8	73
African American	16	52

a. Calculate phi for these data.

b. Calculate Goodman and Kruskal's tau for these data.

c. Using the values that you calculated for phi and tau, how strongly related are the race of the offender and receiving a death sentence?

13.2 Silver Bullet Treatment Services claims to have an effective system for treating criminal offenders. As evidence for the effectiveness of its program, a spokesperson from the organization presents information on rearrest within one year for 100 individuals randomly assigned to the treatment program and for 100 individuals randomly assigned to a control group. The distribution of cases follows:

Experimental Condition	Not Rearrested	Rearrested
Treatment group	75	25
Control group	40	60

a. Calculate phi for these data.

b. Calculate Goodman and Kruskal's tau for these data.

c. Calculate lambda for these data.

d. Based on these three measures of association, what can you conclude about the strength of the relationship between the treatment and rearrest?

13.3 A graduate student is interested in the relationship between the gender of a violent crime victim and the victim's relationship to the offender. To study this relationship, the student analyzes survey data collected on a random sample of adults. Among those persons who had been victims of violent crimes, the student finds the following distribution of cases by gender:

	Relationship of Offender to Victim		
Gender	**Stranger**	**Friend**	**Partner**
Male	96	84	21
Female	55	61	103

a. Calculate V for these data.

b. Calculate Goodman and Kruskal's tau for these data.

c. Calculate lambda for these data.

d. Based on these three measures of association, what can you conclude about the strength of the relationship between gender and the victim's relationship to a violent offender?

13.4 In an attempt to explore the relationship between type of legal representation and method of case disposition, a student working on a research project randomly selects a small sample of cases from the local court. The student finds the following distribution of cases:

	Method of Case Disposition		
Type of Legal Representation	**Convicted by Trial**	**Convicted by Guilty Plea**	**Acquitted**
Privately retained	10	6	4
Public defender	3	17	2
Legal aid	3	13	1

a. Calculate V for these data.

b. Calculate Goodman and Kruskal's tau for these data.

c. Calculate lambda for these data.

d. Based on these three measures of association, what should the student conclude about the relationship between type of legal representation and method of case disposition?

13.5 A researcher interested in the link between attacking other students and being bullied by other students at school used data from a self-report survey administered to a random sample of teenagers. The distribution of responses was as follows:

	Attacked Another Student		
Bullied	**Never**	**Once**	**Two or More Times**
Never	59	22	19
Once	31	44	52
Two or more times	25	29	61

a. Calculate gamma for these data.

b. Calculate τ_b for these data.

c. Calculate Somers' *d* for these data.

d. Interpret each of the three measures of association. What can you conclude about the relationship between being bullied and attacking other students?

13.6 In response to an increasing reluctance of individuals to serve on juries, a study is commissioned to investigate what might account for the public's change of heart. Wondering whether prior jury experience has any effect on how favorably the jury system is viewed, a researcher constructs the following table:

	"How would you rate the current jury system?"			
Served on a jury	**Very Unfavorable**	**Unfavorable**	**Favorable**	**Very Favorable**
Never	22	20	21	26
Once	11	19	12	13
Two or three times	18	23	9	6
Four or more times	21	15	7	4

a. Calculate gamma for these data.

b. Calculate τ_b for these data.

c. Calculate Somers' *d* for these data.

d. Interpret each of the three measures of association. What can you conclude about the relationship between serving on a jury and attitudes about the jury system?

13.7 A researcher interested in the relationship between attitudes about school and drug use analyzed data from a delinquency survey administered to a random sample of high school youth. The researcher was

particularly interested in how well the youth liked school and their use of marijuana. A cross-tabulation of responses revealed the following distribution of cases:

	Smoked Marijuana in the Last Year		
I Like School	**Never**	**Once or Twice**	**Three or More Times**
Strongly agree	52	20	12
Agree	48	26	20
Disagree	31	32	33
Strongly disagree	35	45	50

a. Calculate gamma for these data.

b. Calculate τ_c for these data. Explain why τ_b is not appropriate for these data.

c. Calculate Somers' *d* for these data.

d. Interpret each of the three measures of association. What can you conclude about the relationship between liking school and smoking marijuana?

13.8 A public opinion poll asked respondents whether punishments for convicted criminals should be made more severe, made less severe, or kept about the same. The respondents were also asked to state whether their political views were liberal, moderate, or conservative. A cross-tabulation of the responses to these two questions shows the following distribution of cases:

	Criminal punishments should be . . .		
Political Views	**More Severe**	**About the Same**	**Less Severe**
Liberal	8	54	79
Moderate	35	41	37
Conservative	66	38	12

a. Calculate gamma for these data.

b. Calculate τ_c for these data. Explain why τ_b is not appropriate for these data.

c. Calculate Somers' *d* for these data.

d. Interpret each of the three measures of association. What can you conclude about the relationship between views about politics and attitudes about criminal punishments?

Computer Exercises

Many of the measures of association discussed in this chapter are available in common statistical packages. There are variations in coverage, however, as we note below. There are also sample files containing examples of syntax for both SPSS (Chapter_13.sps) and Stata (Chapter_13.do).

SPSS

Each measure of association discussed in this chapter is available in SPSS with the CROSSTABS command discussed in Chapter 9. Recall from that discussion that the computation of the chi-square statistic was obtained through the /STATISTICS= option. To obtain all the nominal and ordinal measures of association we have discussed in this chapter, we would simply add to the list of association measures:

```
CROSSTABS
/TABLES = row_variable BY column_variable
/STATISTICS = CHISQ PHI LAMBDA GAMMA D BTAU CTAU.
```

where CHISQ is the chi-square, D is Somers' d, and BTAU and CTAU are Kendall's Tau-b and Tau-c, respectively. PHI, LAMBDA, and GAMMA are self-explanatory. Although it is not listed in the command line, Goodman and Kruskal's tau is obtained with the LAMBDA option. Note that you will not need all of these measures for every comparison, and you should pay some attention to the level of measurement and select only those measures that make sense for your data.

In the output generated by this command, you will be presented with the cross-tabulation of the two variables, followed by additional tables that give the various measures of association. Depending on which measures you have requested, you may have three measures of lambda and two measures of Goodman and Kruskal's tau reported. The key to reading the correct values for lambda and tau is to know which variable is the dependent (or outcome) variable.

To illustrate this process, enter the data from Table 13.5 on race and cell block (follow the same process used in the exercises in Chapter 9). Recall in the discussion of the cell block assignment data that we treated cell block as the dependent variable. The value reported for lambda in the line for cell block as the dependent variable will match the value reported in the text. The value reported for Goodman and Kruskal's tau in the line for cell block as the dependent variable will differ slightly from that reported in the text because SPSS does not round the prediction errors to the nearest integer; instead, it records prediction errors with digits after the decimal.

Stata

Measures of association are available through the use of the **tabulate** command by adding either a list of association measures desired or using **all**. Unfortunately, Stata does not compute a wide range of association measures for nominal and ordinal measures.

The command for computing all of the possible (in Stata) measures of association is

tabulate row_variable column_variable, **all**

The output will contain chi-square, Cramer's *V*, gamma, and Kendall's Tau-b. Cramer's *V* will apply to nominal data, while gamma and Kendall's Tau-b will apply to ordinal data.

Problems

The first four problems are likely done more effectively in SPSS, since Stata has limited abilities to compute these measures of association.

1. Enter the data from Table 13.7 into SPSS. Compute the values of Cramer's *V*, tau, and lambda for these data. How do the values of these measures of association compare to those reported in the text?

2. Enter the data from Table 13.13 into SPSS. Compute the values of gamma, τ_c, and Somers' *d* for these data. How do the values of these measures of association compare to those reported in the text? Test the statistical significance of each of the measures of association.

3. Enter the data from Exercise 13.2 into SPSS. Compute the values of phi, tau, and lambda. How do these measures of association compare to the values that you calculated for this exercise? Test the statistical significance of each of the measures of association.

4. Enter the data from Exercise 13.6 into SPSS. Compute the values of gamma, τ_c, and Somers' *d* for these data. How do these measures of association compare to the values that you calculated for this exercise?

5. Open the NYS data file (nys_1.sav, nys_1_student.sav, or nys_1.dta). Each pair of variables listed below was tested for a relationship using the chi-square test in the computer exercises at the end of Chapter 9. For each pair of variables, determine the level of measurement (nominal or ordinal) and the dependent and the independent variables; then compute appropriate measures of association to the extent that you are able. Interpret each of the measures of association that you have computed. Test the statistical significance of each of the measures of association. What can you conclude about the relationship between each pair of variables?

   a. What is the relationship between ethnicity and grade point average?

   b. What is the relationship between marijuana use among friends and the youth's attitudes about marijuana use?

c. What is the relationship between the importance of going to college and the importance of having a job?

d. What is the relationship between grade point average and the importance of having a job?

e. What is the relationship between the youth's sex and the importance of having friends?

f. What is the relationship between the importance of having a job and the youth's attitudes about having a job?

6. Open the Pennsylvania Sentencing data file (pcs_98.sav or pcs_98.dta). Each pair of variables listed below was tested for a relationship using the chi-square test in the computer exercises at the end of Chapter 9. For each pair of variables, determine the level of measurement (nominal or ordinal) and the dependent and the independent variables; then compute appropriate measures of association. Interpret each of the measures of association that you have computed. Test the statistical significance of each of the measures of association. What can you conclude about the relationship between each pair of variables?

   a. Is the sex of the offender related to the method of conviction?

   b. Is the race–ethnicity of the offender related to whether the offender was incarcerated or not?

   c. Is the method of conviction related to the type of punishment received?

   d. Is the type of conviction offense related to the method of conviction?