8 Measures of Association: Ordinal and Nominal Variables

Researchers in the social and the behavioral sciences generally study variables measured at the interval level of measurement. It is also possible for a variable to be measured at either the nominal or ordinal level of measurement. (The topic of level of measurement was presented in Chapter 1.) For instance, the variable of political party is at the nominal level of measurement. In this case each person does not have a score or number but rather each person is a member of a category. A variable, such as political party, that is categorical and not numeric is called a *nominal variable*. For a nominal variable each person is a member of one discrete category as opposed to each person receiving a numeric score. A nominal variable with only two categories is called a *dichotomy*.

Ordinal variables are also encountered in the social and behavioral sciences. Variables measured at the ordinal level of measurement permit a rank ordering of the objects. One can make a case that many variables in the social and behavioral sciences are measured at the ordinal level.

Although not as common as interval variables, nominal and ordinal variables are hardly unusual. Many of the standard demographic variables are nominal. For instance, gender, ethnicity, religion, and geographic region of residence are all nominal variables. Many medical and physical variables are also nominal: eye color, blood type, left- versus righthandedness, and diagnostic category. Many of the responses that are studied in research are nominal in nature. Whether a person agrees or disagrees on a survey, whom the voter prefers in an election, what product a customer purchases, whether a

surgical procedure results in life or death, whether a subject recalls a nonsense syllable, and whether a nerve cell fires when stimulated are all examples of nominal variables. Nominal variables are quite common in research.

Ordinal variables can be established if the measurements are a rank ordering. Some variables, such as birth order and military rank, are clearly at the ordinal level of measurement. The boundary between the ordinal and the interval level of measurement is quite cloudy. The determination is often a matter of preference.

The previous six chapters focused on the various descriptive statistics for variables measured at the interval level of measurement. In this chapter various descriptive statistics for nominal and ordinal variables are presented. Most of the discussion concerns measures of association for these variables.

Shape, Location, Variation, and Transformation

This section concerns the issues of shape, location, and variation for both nominal and ordinal variables. First considered are nominal variables.

Nominal Variables

The distribution of a nominal variable is simply a set of frequencies. For instance, of 132 people polled, 28 people intend to vote Democratic, 36 Republican, and 68 intend not to vote at all. The numbers 28, 36, and 68 are frequencies. Or a particular surgical procedure resulted in 42 deaths and 778 lives saved. Or 91 nerve cells fired and only 3 did not. The distribution of a nominal variable is the number of observations in each category.

The categories of a nominal variable are not numeric. So when the numbers are graphed in a histogram, the bottom axis or X axis is not numeric. The resulting distribution is called a *bar graph*. In a bar graph, unlike a histogram, the bars are separated by a space. Because the categories are not numeric, the ordering of the categories on the X axis is arbitrary. One still should compute the relative frequencies for each category. For instance, if 58 respondents say yes and 21 say no, one would compute $58/(58 + 21) = .73$ for yes and $21/(58 + 21) = .27$ for no.

What then can be said about the shape of the distribution? With a nominal variable, one can describe how flat the distribution is. A flat distribution is one in which each of the categories is equally likely. A dichotomy in which the categories are quite uneven (say 75% to 25%) is said to be skewed.

Because the categories of a nominal scale are not ordered, a median makes

no sense; and because they are not quantitatively ordered, a mean likewise does not represent a meaningful measure of central tendency. However, the mode can be determined. The modal category has the largest frequency. Hence for the nominal scale of political party, the party most frequently chosen would be the mode.

The formulas for variability presented in Chapter 4 are not appropriate for categorical variables. However, there is relatively more variation if the distribution is flat. Nearly equal numbers of persons in the categories results in greater variation. If the number of persons in one category is relatively large, then there is relatively little variation.

All of the data transformations discussed in Chapter 5 involve a quantitative operation on the data. Because with nominal variables there are only categories and not numbers, all of those transformations are inappropriate. Nonetheless, nominal variables are in a sense transformed when categories are collapsed. For instance, in surveys it is common practice to treat "don't know" and "no opinion" as the same response. There are two helpful rules for determining which categories to collapse. First, one should consider as good candidates for collapsing those categories whose occurrences are infrequent, say less than 5%. The second rule is to combine categories that are conceptually similar. For instance, if the categories are white, black, and Hispanic, it may be sensible to collapse black and Hispanic to form a minority group category.

Recall that the definition of a nominal variable is one in which each observation is placed in one and only one category. The categories of a nominal variable are said to be mutually exclusive (no person may be in two categories) and exhaustive (each person must be in at least one category). Occasionally, a given nominal variable may violate these rules, but the violations can be easily remedied. For instance, there may be some persons who are neither Democrat, Republican, nor Independent, and, less likely, there could be a person who claims to be a member of two parties. Those persons who do not fall into any category can be put into a residual category of "other." Alternatively, if there are few persons who fit in no category (less than 5%), they can be dropped from the sample. For those who are members of two categories, a new category of "both" could be created or again if they are few in number they could be dropped from the sample.

Occasionally, it is useful to treat a nominal variable *as if* it were a numeric variable. In this case, a researcher arbitrarily assigns numbers to the various categories of the nominal variable. For instance, for the variable of gender, men may be given a score of zero and women a score of one. This is said to be an arbitrary assignment because the researcher could have just as easily given men a score of one and women a score of zero, or men a score of 50 and women a score of –38. The arbitrary assigning of numbers to levels of a nominal variable is called creating a *dummy variable*. Another example of dummy variables is as follows:

assign a 1 for Catholics,
assign a 2 for Protestants,
assign a 3 for Jews, and
assign a 4 for others.

Dummy variables are used in computing the phi coefficient, which is a measure of relationship between two dichotomies that is presented later in this chapter.

For a dichotomy, the usual convention is to assign a one to one category and a zero to the other. The mean of such a dummy variable is the proportion of persons assigned a one. The standard deviation[1] is the square root of $n/(n-1)$ times the product of the proportion assigned a one and one minus that proportion. So if p is the proportion of people assigned a one, then the mean of the dummy variable is p and the standard deviation is the square root of $np(1-p)/(n-1)$.

Ordinal Variables

If a variable is truly measured at the ordinal level of measurement, then its shape contains nothing theoretically interesting. The numbers reveal only the relative position of the persons in the sample and nothing about the distance between two scores. Shape then is virtually meaningless for a variable that is at the ordinal level of measurement. However, if an ordinal variable has relatively few levels and there are many observations, there are many tied observations. Two observations are tied if the researcher is unable to determine which observation has more of the quantity that is measured. When an ordinal variable has many tied observations and few levels, a bar graph can be drawn. The ordinal variable would be on the X axis and the number of tied observations would be on the Y axis.

The concepts of central tendency and variability have no meaning for ordinal variables. However, the median, though not quantitatively interpretable, can be of interest. For instance, if it is known that, in terms of the continental United States, South Carolina is the state with the median population, then it is known what state it is that ranks in the middle.

If a variable is measured at the ordinal level, it is a common practice to transform the scores by rank ordering them (see Chapters 5 and 18). For reasons discussed in Chapter 18, the mean and the standard deviation of these ranks are of statistical interest. Given a sample of n scores, their mean rank must be

$$\bar{X} \text{ of ranks} = \frac{n+1}{2}$$

[1]This formula is usually presented as $p(1-p)$. However, because this text uses $n-1$ in the denominator for variance, the formula in the text is appropriate.

Given no ties, the standard deviation[2] of ranks is:

$$\text{standard deviation of ranks} = \sqrt{\frac{n^2 + n}{12}}$$

So if $n = 10$, then the mean is 5.5 and the standard deviation is 3.03. If there are ties, the usual formulas for standard deviation must be employed.

Relationship

The remainder of this chapter concerns the measures of association between variables that are either at the nominal or the ordinal level of measurement. Considered first is the relationship between two dichotomous variables. A dichotomous variable is a nominal variable with only two categories. Next is discussed the general problem of measuring the association between two nominal variables. Finally, the issue of how to measure the association between two variables measured at the ordinal level of measurement is discussed.

Two Dichotomies

Of key interest is whether there is any relationship between two nominal variables. Is political party related to voting behavior? Is a surgical procedure related to survival? Are people more likely to give blood depending on the type of appeal? All of these questions are concerned with the relationship between two nominal variables.

In Table 8.1 is a table of numbers. The data are taken from a study by Korytnyk and Perkins (1983). They placed 29 male, heavy drinkers in a situation in which the subjects could write graffiti on a wall. Of the subjects, 15 were given tonic water and 14 were given the equivalent of two drinks of alcohol. The two variables being associated are beverage consumed (tonic versus alcohol) and whether the subject wrote graffiti or not. Beverage makes up the rows of the table and graffiti behavior, no and yes, makes up the columns. The numbers in the table are called counts. There are four counts. For instance, there are 14 who received tonic and did not write on the walls, and 7 alcohol drinkers who wrote on the wall.

The entries in a table, called frequencies, are the number of persons in that cell. Because each variable is made up of two categories, the table is called a 2 by 2, or 2 × 2, table.

Again the rows of the table are beverage (tonic or alcohol) and the columns

[2]The numerator of this formula is often presented as $n^2 - 1$. However, because this text always uses $n - 1$ as the denominator for variance, $n^2 + n$ is appropriate.

TABLE 8.1 **A 2 × 2 Table**

		Graffiti: No	Graffiti: Yes	
Beverage	Tonic	14	1	15
	Alcohol	7	7	14
		21	8	29

are graffiti (no or yes). It is a common practice to add the counts across both rows and columns. There are 15 tonic and 14 alcohol drinkers, and 21 who did not write on the wall and 8 who did. These sums are commonly called the *margins*. The column margins are 21 and 8. The sum of the row margins (15 + 14) should equal the sum of the column margins (21 + 8), and this provides a useful computational check. The total sum is written in the bottom right-hand corner.

Of special interest is whether there is any association between beverage consumed and graffiti behavior. Stated differently, the question is whether those who drink alcohol are more or less likely to vandalize than those who do not. This question concerns whether the two nominal variables are associated. To measure association the researcher can chose among the percentage difference, the phi coefficient, or the logit difference.

Percentage Difference. The simplest and perhaps most natural measure of association is to compute the percentage of tonic drinkers who write on the wall and the percentage of alcohol drinkers who write on the wall. Using the data in Table 8.1, only 6.7% (1/15) of tonic drinkers write on the wall, and 50.0% (7/14) of the alcohol drinkers write on the wall. This difference between 50.0% and 6.7% is 43.3%. This is the percentage difference measure of association for nominal data. The percentage could be computed for each column. That is, of those who do not write on the wall, 66.7% (14/21) are tonic drinkers, and of those who do write on the wall, 12.5% (1/8) are tonic drinkers. This difference is 54.2%. As this example shows, the percentage difference measure is not necessarily a symmetric measure of association. The percentage difference measure may change if the percentages are calculated across rows or across columns.

The percentage difference measure can be viewed as a regression coefficient. One variable is denoted as the predictor variable and is dummy-coded

zero and one. The other variable is the criterion and is dummy-coded zero and 100. If a regression coefficient were computed for these two dummy variables, its value would be identical to the percentage difference measure.

Phi Coefficient. The second measure is phi, which is symbolized by ϕ. This measure of association is not as commonly used as percentage difference. Phi is a correlation coefficient. So if there is no relationship, phi is near zero, and if there is a near-perfect relationship phi is near 1.00 or –1.00. To understand how phi is computed, consider Table 8.2. The two nominal variables have been designated X and Y. The four frequencies are designated a, b, c, and d. The phi coefficient is found as follows:

$$\phi = \frac{ad - bc}{\sqrt{(a + b)(c + d)(a + c)(b + d)}}$$

For example, in Table 8.1 phi equals .484. As was stated earlier, the phi coefficient is a correlation coefficient. In Table 8.2 there are dummy variables for both variables. The first row is given a 1 and the second a 0. The first column is given a 1 and the second column a 0. Using these dummy variables, one can compute the correlation between the two variables. This correlation equals phi. Ordinarily, it is much simpler to use the formula presented earlier. As will be seen in Chapter 17, phi is a useful number for computing other statistics. Unlike the percent difference measure but like the correlation coefficient, the measure phi is symmetric.

In Chapter 7, it was stated that r equals the square root of the product of the two regression coefficients. This fact can be used to relate phi to the two percentage difference measures: Phi times 100 equals the square root of the product of the two percentage difference measures. So, for the example, 100 times phi (48.4) equals, within rounding error, the square root of the product of the two percentage difference measures (43.3 × 54.2).

TABLE 8.2 **Symbols for a 2 × 2 Table**

		Variable Y: 1	Variable Y: 0	
Variable X	1	a	b	$a + b$
	0	c	d	$c + d$
		$a + c$	$b + d$	n

Logit Difference. The percentage difference is based on the regression coefficient and phi is based on the correlation coefficient. A third measure of association in a 2 × 2 table is the logit difference. In Chapter 5, the logit transformation of proportions is presented. A logit of a proportion p is

$$\ln\left(\frac{p}{1-p}\right)$$

where ln is the logarithm to base e. In Appendix A in the back of the book is a table for the conversion of a proportion into a logit. As stated in Chapter 5, the logit of zero corresponds to a proportion of .5, a positive logit to a proportion greater than .5, and a negative logit to a proportion less than .5. Using the symbols in Table 8.2, the logit for the upper row is $\ln(a/b)$ and for the lower row is $\ln(c/d)$. The logit difference is then

$$\ln\left(\frac{a}{b}\right) - \ln\left(\frac{c}{d}\right)$$

Although it is not intuitively obvious, the logit difference is a symmetric measure. That is, it is true that

$$\ln\left(\frac{a}{b}\right) - \ln\left(\frac{c}{d}\right) = \ln\left(\frac{a}{c}\right) - \ln\left(\frac{b}{d}\right)$$

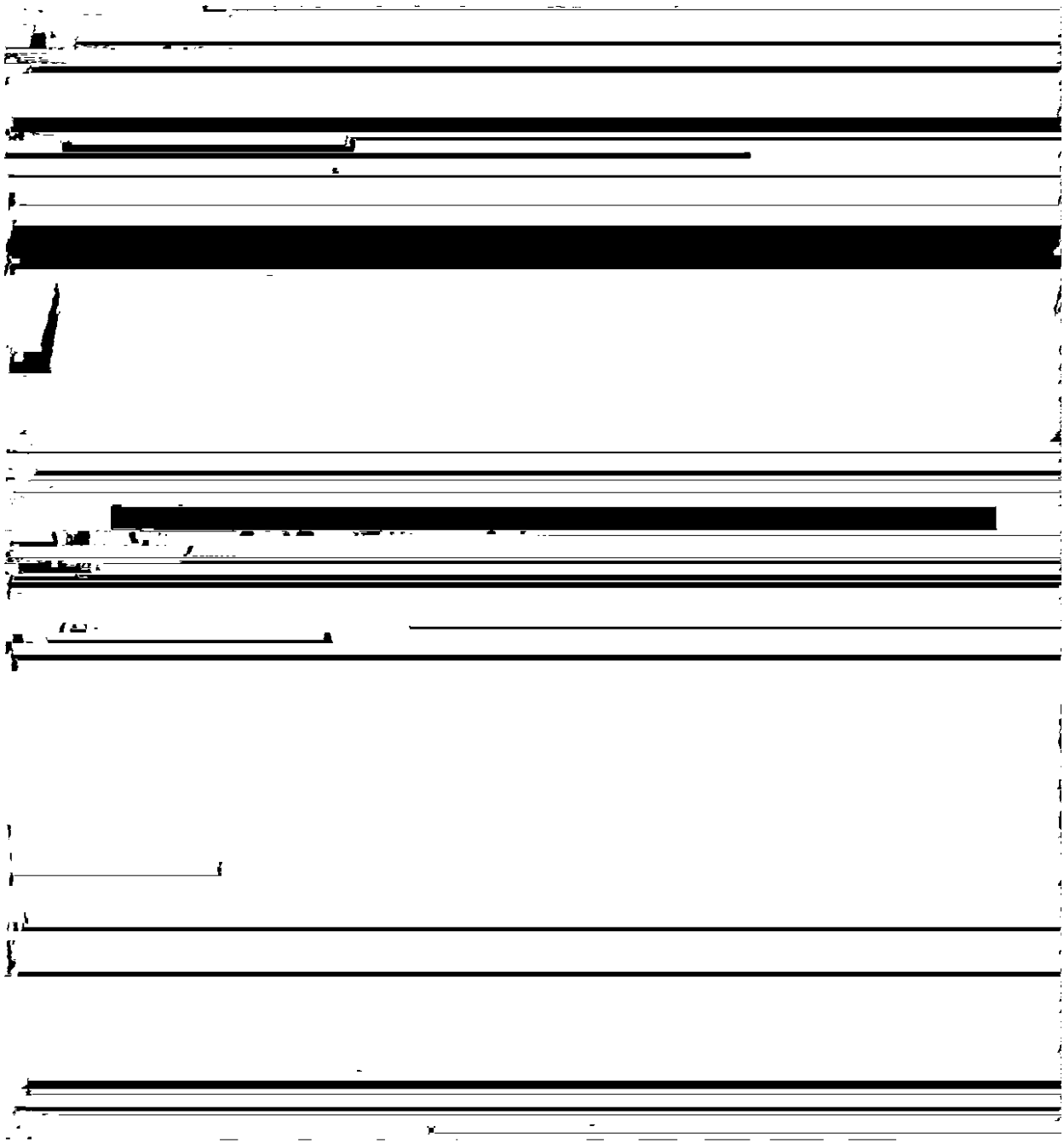
This is true because each equals

$$\ln\left[\frac{a/b}{c/d}\right]$$

The term in the parentheses is a ratio of two odds and is, therefore, called the *odds ratio*. For the example, it is the odds of tonic drinkers writing on the walls divided by the odds of alcohol drinkers doing so. Thus, the logit difference can be interpreted as the natural logarithm of the odds ratio. This odds ratio formula for the logit difference is simpler than the logit difference formula, because the odds ratio formula involves taking a logarithm only once. The logit difference for the example equals 2.64. The odds ratio is 14 and its natural logarithm is 2.64, which is equivalent to the logit difference measure.

If any of the frequencies equals zero, the logit difference measure is not defined. To remedy this problem, .5 is added to each of the frequencies before the logit difference is computed.

Interpretation and Comparison. Table 8.3 gives formulas for the percentage difference, phi coefficient, and logit difference using the symbols presented in Table 8.2. Although the measures are quite different, when one of them equals zero, the other two also equal zero. They are all zero only when $ad = bc$ (see Table 8.2).



Percentage Difference

$$100\left[\frac{a}{a+b} - \frac{c}{c+d}\right]$$

Phi

$$\frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

Logit Difference

$$\ln\left(\frac{a}{b}\right) - \ln\left(\frac{c}{d}\right)$$

TABLE 8.4 **Logit Difference in Terms of Percentage Difference Measure in which the Margins Are 50/50**

Logit Difference	Percentage Difference
.00	0.0
.05	1.2
.10	2.5
.15	3.7
.20	5.0
.25	6.2
.30	7.5
.35	8.7
.40	10.0
.45	11.2
.50	12.4
.60	14.9
.70	17.3
.80	19.7
.90	22.1
1.00	24.5
1.25	30.3
1.50	35.8
1.75	41.2
2.00	46.2
2.50	55.4
3.00	63.5
5.00	84.8

difference or phi to be as large as it is for the group in Table 8.5. Only for the logit difference can the measure of association possibly remain stable. It is for this reason that many researchers prefer the logit difference measure. The logit difference measure tends to replicate better across time and settings. The logit difference measure tends not to be as affected by changes in the margins as the percentage difference and phi are.

What is the most appropriate way to measure association in a 2 × 2 table? The answer depends on the purposes of the researcher. If a measure is desired that is simple to compute and easy to interpret, then the percentage difference measure is probably best. If the variables cannot be distinguished as a predictor and criterion, then phi is probably best. If the measure is to be computed for different samples with different margins, then the logit difference is probably best.

Nominal Variables with More than Two Levels

The measures of association between two nominal variables, at least one of which has more than two levels, are analogous to the measures of association

TABLE 8.5 **Illustration of the Generalizability of the Logit Difference**

	Cause of Death: Lung Cancer	Cause of Death: Other	
Smoker	681	7831	8512
Nonsmoker	123	8391	8514
	804	16222	17026

Percentage Difference: 6.56
Phi: .155
Logit Difference: 1.780

between two dichotomous variables. However, a complete presentation of these techniques is beyond the scope of this book. The reader is referred to more advanced texts that present these measures (Fienberg, 1977; Reynolds, 1977).

One can compute percentages across either rows or columns. So to compute the percentage for each row, an entry is divided by its row total. The sum of the percentages across each row should add to 100. As an example, consider the data in Table 8.6. The data are taken from a vote in the United States Congress in 1836. The issue concerned a matter related to slavery, a vote for the law being proslavery vote. There were three possible vote alternatives: yes, abstain, and no. The 225 congressmen are classified according to the section of the country that they represented. So 61 congressmen from the North voted yes. Beneath each number, in parentheses, is the percentage computed across rows. For instance, the percentage of congressmen voting no from the North is 60 divided by 133 (the row margin) times 100 or 45%. The table clearly shows that support for the proslavery position was located in the South and border states.

Ordinal Variables: Spearman's Rho

Sometimes the researcher might question whether a given variable is measured at the interval level of measurement. The researcher may believe that the variables are measured at only the ordinal level of measurement. That is, the numbers indicate only the relative positions of persons and not any quantitative difference. The association of variables measured at the ordinal level of measurement can be measured by Spearman's rho. Although other measures,

TABLE 8.6 **Frequencies for a 3 × 3 Table of Vote by Region and Row Percentages in Parentheses**

Region Represented	Vote: Yes	Abstain	No	Totals
North	61 (46)	12 (9)	60 (45)	133
Border	17 (71)	6 (25)	1 (4)	24
South	39 (57)	22 (32)	7 (10)	68
Totals	117	40	68	225

Data were taken from Benson and Oslick (1969).

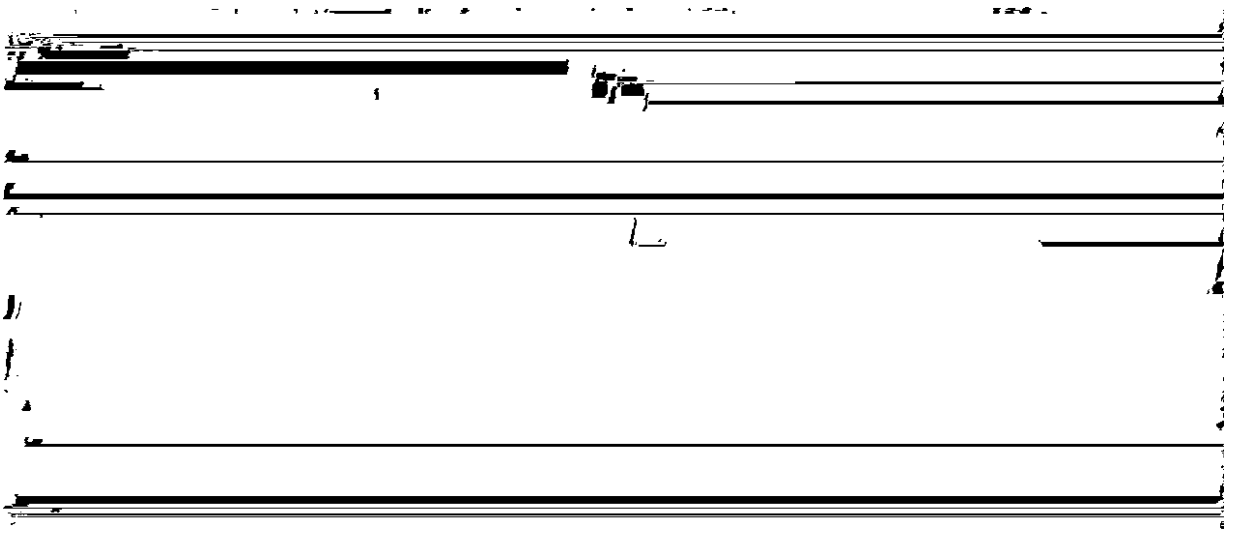
such as Kendall's tau and Goodman and Kruskal's gamma, are also employed, rho is by far the most common measure of ordinal association.

Variables measured at the ordinal level of measurement can be transformed by a rank-order transformation. This transformation is described in Chapter 5. Each score is given a rank from one to n. If two or more scores are tied, they are each given the average rank. Then these rank-order scores can be correlated using the formula for the correlation coefficient presented in the previous chapter. A correlation coefficient of rank orders is called *Spearman's rho* or the *rank-order correlation* and is denoted as r_S. Fortunately all the computational work of correlating the two sets of ranks can be avoided. There is a computational shortcut. It involves first computing the difference between each pair of ranks, D_i. It happens that

$$\text{Spearman's rho} = 1 - \frac{6\sum D_i^2}{n(n^2 - 1)}$$

where n is sample size and D_i is the difference in ranks for the ith pair of scores. This formula presumes no ties in the ranks. If there are ties, one must use the formula for the correlation coefficient given in Chapter 7.

The six in the formula for Spearman's rank-order coefficient strikes some as odd. Its presence is due to the formula presented earlier in the chapter for the standard deviation of ranks. The denominator of that formula has a twelve in it which brings about the six in Spearman's rho or r_S.



Another use of Spearman's rho is to control for nonlinearity. The standard measures of association discussed in the previous two chapters presume that the relationship is linear. If the relationship between variables is nonlinear but does not change direction (see Chapter 7), then Spearman's rank-order coefficient is a useful measure because linearity is not assumed.

In 1884, Francis Galton measured the strength of more than 9000 persons who visited various museums in London. Johnson, McClean, Yuen, Nagoshi, Ahern, and Cole (1985) report the average degree of hand strength that Galton obtained for men from 11 through 25 years of age:

Age	*Hand Strength*	*Age*	*Hand Strength*
11	33.11	19	80.07
12	37.53	20	80.19
13	40.12	21	81.12
14	48.68	22	80.46
15	57.99	23	79.86
16	67.41	24	81.36
17	73.94	25	82.27
18	78.25		

The scores are rank ordered by age in ascending order and the differences (D) and the squared differences (D^2) are computed:

Age	*Hand Strength*	D	D^2
1	1	0	0
2	2	0	0
3	3	0	0
4	4	0	0
5	5	0	0
6	6	0	0
7	7	0	0
8	8	0	0
9	10	−1	1
10	11	−1	1

Summary

Nominal and ordinal variables are commonly used in the social and the behavioral sciences. Because the standard statistics that were developed in the previous chapters are for variables that are measured at the interval level of measurement, new procedures are developed. Because the variables are not at the interval level of measurement, the usual measure of central tendency, the mean is not appropriate. However, the mode can be used as a measure of central tendency for nominal variables and the median for ordinal variables. For neither the nominal nor ordinal levels of measurement does the shape of distribution make much sense, except for plotting the frequencies of a nominal variable in a bar graph.

For nominal variables it is sometimes necessary to collapse or eliminate categories. When numbers are assigned to levels of a nominal variable, the resulting variable is called a *dummy variable*.

There are three major measures of association between two nominal variables. They are the *percentage difference*, the *phi coefficient*, and the *logit difference*. The percentage difference is the simplest measure of association. One variable is denoted as a predictor and the other as the criterion. The percentages are computed for those responding in one category of the criterion for each of the categories of the predictor variable. The percentage difference is the difference between these two percentages. The phi coefficient is a correlation coefficient between dummy variables. For the logit difference, the odds of responding are computed for each category. These odds are logged and then differenced. The logit difference, while more difficult to interpret, is more likely to generalize across different samples. Only the percentage difference measure can be easily generalized for nominal variables with more than two levels.

One standard measure of association between two ordinal variables is *Spearman's rank-order correlation*, also called *Spearman's rho*. Spearman's rank-order correlation r_S is a correlation coefficient between ranks. This measure is based on the difference between the ranks of the two variables. The rank-order correlation can be used to measure the association between variables measured at the interval level of measurement when nonlinearity is suspected.

Problems

1. For the following sets of categories, by collapsing categories, create a new nominal variable with only two categories.

   a. Protestant, agnostic, atheist, Catholic, Jewish
   b. rainy, clear, cloudy, snowy
   c. radio, television, stereo, tape deck
   d. anger, disgust, happiness, fear

2. a. Imagine that 26 persons agree and 49 disagree with a particular statement. If responses are dummy coded (1 = agree, 0 = disagree), what is the mean and standard deviation of the dummy variable?
   b. If 32 agree and 42 disagree, what would be the mean and standard deviation?

3. If 76 Democrats favor capital punishment and 73 disapprove, and 108 Republicans approve and 111 disapprove, set up the 2 × 2 table with margins. Compute phi, the percentage difference (treating political party as the predictor variable), and the logit difference.

4. For the following table compute the percentage difference (treating gender as the predictor variable), phi coefficient, and logit difference. The column variable is whether the person is a smoker or not. Interpret each measure.

	Smoking: Yes	Smoking: No
Women	71	28
Men	19	48

5. For the following table compute the percentage difference (treating age as the predictor variable), phi coefficient, and logit difference. Interpret each measure.

	Yes	No
Over 30	28	83
Under 30	16	9

6. Taylor and Ferguson (1980) asked 200 students where they went when they wanted to be alone (solitude) and when they wanted to talk with a close friend (intimacy). Their answers were then coded as one of three kinds of territory: primary, where the individual can control access to the area; secondary, where the control of access is shared with others; and public, where the individual has no control over access. The table below gives the number of responses in each category.

	Intimacy		
Solitude	*Primary*	*Secondary*	*Public*
Primary	24	2	35
Secondary	3	0	4
Public	59	8	65

a. Compute the percentage of students who chose each kind of territory for intimacy. (Compute the percentages for each column.)
b. What proportion chose each kind of territory for solitude? (Compute the percentages for each row.)

7. In a study of privacy regulation, Vinsel, Brown, Altman, and Foss (1980) had freshman dormitory residents check a list of techniques they might have used to avoid contact with others. A year later, 19 of the students had left the university (dropouts), while 54 were still enrolled (stay-ins). They found that five of the dropouts and nine of the stay-ins used loud music to avoid contact.

   a. Create a 2 × 2 table for the use of music (yes or no) and enrollment status (dropout or stay-in).
   b. For the table compute the percentage difference (treating enrollment status as the predictor variable), phi coefficient, and logit difference. Interpret each measure.

8. The following table shows the number of hurricanes that occurred in the years 1886–1981.

Month	*Hurricanes*
January	0
February	0
March	0
April	1
May	3
June	21
July	32
August	135
September	176
October	85
November	18
December	2

   Treat the twelve months as categories and construct a bar graph of the data.

9. Harrison (1984) asked dormitory residents who had chosen or been assigned to their dorms whether they wanted to change roommates at the end of the semester. The following table gives the results. Compute the percentage difference (treating choice vs. assignment as the predictor variable), the phi coefficient, and the logit difference. Interpret each measure.

Dorm	Change Roommate: No	Change Roommate: Yes
Chose	22	3
Assigned	34	12

10. Compute Spearman's rank-order correlation for the following set of scores of seven persons on variables X and Y.

Person	X	Y
1	12	9
2	15	7
3	13	6
4	9	5
5	6	8
6	14	4
7	19	1

11. Below are the selected life expectancies for seven countries for males and females. Compute Spearman's rank-order correlation r_S for the countries:

	Life Expectancy (years)	
Country	*Males*	*Females*
India	51	50
Indonesia	45	48
Brazil	58	63
Bangladesh	50	47
Nigeria	40	43
U.S.A.	70	78
Japan	73	79

Interpret the value of the rank-order correlation.

12. For the numbers in problem 11 compute the ordinary correlation r between the life expectancy for males and females in seven countries. Why are r and r_S different?