17 Models for Nominal Dependent Variables

The preceding five chapters discussed models in which the dependent variable is assumed to be measured at the interval level of measurement. In this chapter models are considered in which the dependent variable is measured at the nominal level of measurement. And in the final chapter the dependent variable is assumed to be measured at the ordinal level of measurement. Techniques that do not assume an interval dependent variable are sometimes referred to as nonparametric or distribution-free methods. The term *distribution-free* is preferred because so-called nonparametric tests do test hypotheses about parameters. Methods that presume normality and homogeneity of variance such as the two-sample t test, analysis of variance, and regression will be called *distribution-tied* methods.

There are three major reasons for employing the methods described in this chapter and the next. The first reason is that sometimes the data are clearly not at the interval level of measurement. The dependent measure may be a set of ranks or a set of categories. In these cases it would be clearly inappropriate to use the methods described in the previous five chapters. So if the level of measurement of the dependent variable is clearly not at the interval level of measurement, the methods presented in this and the next chapter are appropriate.

The second use of distribution-free procedures is that one may be reasonably confident that the dependent variable is at the interval level of measurement, but one is worried about the assumptions made to perform a t or F test. In particular, one may be especially concerned that the assumption of a normal distribution for the residual variable is false. The dependent variable may be highly skewed or bimodal, and so it is quite likely that the residual variable does not have a normal distribution. One then desires to do a statistical test, but one is unwilling to make assumptions concerning the distribution of the residual variable. Because distribution-free methods make no assumptions concerning the distribution of any of the variables, they can

be used with bimodal or highly skewed distributions. This is why these methods are called distribution-free.

If the residual variable does have a normal distribution and all the other assumptions are met, there is a cost in not doing an analysis that presumes the interval level of measurement. The techniques described in this and the next chapter have less power than the procedures described in the previous chapters when the assumptions made by distribution-tied tests are true. Thus analysis of variance and the two-sample t test are more powerful statistical procedures than the distribution-free procedures described in this and the next chapter. However, if the classical assumptions of normality and homogeneity of residuals do not hold, the p values obtained from analysis of variance are not correct and are usually too liberal, resulting in too many Type I errors. It can even happen that for some distributions, a distribution-free method is more powerful than a distribution-tied method.

A distribution-free test is ordinarily less powerful than its distribution-tied cousin because the distribution-free test ignores the interval information in the data. Consider the following pattern of numbers of two samples A and B.

A: 1, 2, 3, 6

B: 7, 8, 9, 12

There is no overlap in the numbers and the means (3.0 for sample A and 9.0 for sample B) differ by six units. Consider the pattern of the following numbers from two samples.

A: 1, 2, 3, 6

B: 107, 108, 109, 112

Again there is no overlap, but now the means (3.0 for sample A and 109.0 for sample B) differ by 106 units. A distribution-free test would see no difference between the two patterns, whereas a distribution-tied method would see the second pattern as more convincing evidence that the two groups differ.

However, distribution-free tests do have their advantages. Distribution-tied tests believe even the most anomalous aspect of the data. Consider again the first pattern of the numbers of two samples A and B:

A: 1, 2, 3, 6

B: 7, 8, 9, 12

There is no overlap in the numbers and the means differ by six units. Consider the following numbers from two samples.

A: 1, 2, 3, 6

B: 7, 8, 9, 120

The numbers are exactly the same except that the last number in the B sample

is ten times larger in the second pattern. The distribution-free test would see no difference between the two patterns, whereas a distribution-tied method would see the second pattern as much more convincing evidence that the two groups differ even though the value of 120 would appear to be an outlier.

The third reason for choosing a distribution-free test is that it tests a different null hypothesis from the null hypothesis tested by the distribution-tied analog. For instance, consider the following two samples.

A: 1, 2, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 7, 8

B: 1, 1, 1, 2, 2, 2, 2, 7, 7, 7, 7, 8, 8, 8

Both groups have means of 4.5, but clearly the groups differ. Sample B has more extreme scores than sample A. A distribution-free test can reveal such a difference, but a t test cannot.

Although there are clear-cut cases in which a distribution-free statistic is clearly superior to its distribution-tied cousin, the choice between the two may be more a matter of taste and custom than of right or wrong. For instance, researchers in medicine are much more likely to employ a distribution-free method than researchers in economics, even though data in medicine are no less likely to be normally distributed than in economics. Perhaps the preference is explainable by need to be somewhat more conservative when lives are at stake than when dollars are. I suspect, however, that the real reason has more to do with custom than anything else.

In cases in which the researcher is in doubt about the type of analysis, both types of tests might be employed. Most of the time the two sets of results agree. In such a case the distribution-tied tests are reported with mention that the distribution-free results are in essential agreement. In cases in which the analyses are in conflict, usually the distribution-free results are reported because they tend to be more conservative.

This chapter considers distribution-free tests in which the dependent variable is at the nominal level of measurement. Two basic types of models are considered. In the first, hypotheses concerning the distribution of a nominal dependent variable are tested. In the second, both the independent and dependent variable are at the nominal level of measurement. For this second model, either the scores can be independent across levels of the independent variable or they can be nonindependent. Different analysis strategies for models are needed when the groups are independent and when they are nonindependent.

First, this chapter shows how to test whether a nominal variable affects a second nominal variable in which the groups are independent. This test is commonly called a χ^2 *test of independence*. Its distribution-tied analogs are the two-group t test, one-way ANOVA, and regression. The second test considered in this chapter is the *McNemar test*, which evaluates the effect of a dichotomous nominal variable on a dichotomous dependent variable in which

the groups are nonindependent. Its distribution-tied analog is the paired t test which was presented in Chapter 13. The final test that is discussed evaluates the adequacy of an a priori prediction of a nominal variable's distribution. The test is commonly referred to as a χ^2 *goodness of fit* test. Its distribution-tied analog is a t test of a constant which is presented in Chapter 12.

As was explained in Chapter 8, for a nominal variable the data can be converted into frequencies. A *frequency* of a category equals the total number of objects for the category of the nominal variable. For the statistical tests presented in this chapter, the χ^2 distribution is the sampling distribution that is employed. In all cases the distribution of the test statistic is approximately χ^2. The test statistic for these χ^2 tests always compares the observed or actual frequencies to those frequencies expected under a restricted model.

Test of Independence of Two Nominal Variables

In this case there are two nominal variables and the issue is whether the two variables are associated. One variable may be distinguished as independent and the other dependent or they may not be. Such a distinction does not affect the p value but it does affect the interpretation of the result.

As an example, consider a study by Brown (1981). He had a pair of persons stand in a mall talking to one another. Persons approaching the pair could either walk through the pair or walk around. Brown varied the racial composition of the pair. They were either both black, both white, or mixed race. So the independent variable is racial composition of the pair, and the dependent variable is the behavior of the subject: walking through versus around. A total of 508 subjects were observed, and the results are shown in Table 17.1.

The first row of the table consists of those who walked through. For instance, a total of 125 persons walked through the black pair. The second row consists of those who walked around. The final row is called the set of column margins and consists of the number of persons in the sample for type

TABLE 17.1 **Observed Frequencies for the Brown (1981) Study**

	Racial Composition			
Behavior	Black	White	Mixed	Total
Through	125	67	65	257
Around	69	76	106	251
Total	194	143	171	508

of pair. The final column contains the row margins. They give the total number of persons who walked through and around. The number in the bottom right-hand corner, 508, is the total number of persons in the study.

As discussed in Chapter 8, a table of frequencies is, by itself, not very interpretable. To increase interpretability the percentage of those who walked through for each racial composition is computed. Percentages are computed for each column because racial composition is the independent variable and behavior is the dependent variable. The result is shown in Table 17.2. The subjects are most likely to walk through the pair when the pair is black and least likely when the pair is mixed. Interestingly, the mixed-pair percentage does not fall halfway between the black and white pairs.

It might be asked whether these results could be explained by sampling error. Is it possible that, by chance, the subjects in the black condition just happened to be persons who would walk through any pair? Can the hypothesis that the racial composition does not affect behavior and that the observed differences are due to sampling error be ruled out?

If there is no association between the two nominal variables, then it is said that the two variables are independent. Thus, the complete model assumes that the two nominal variables are associated and the restricted model is that the two variables are independent.

To evaluate the restricted model, it is necessary to estimate the number of subjects who would walk through the black pair if the variables of racial composition and behavior were independent. The actual or observed number is compared with the *frequency expected* if the two variables were independent.

The expected frequency for a given cell equals the row margin times the column margin divided by the total number of persons. (Note that *frequencies,* not proportions, are used.) So, the expected number of persons who walk through the black pair is the row margin (257) times the column margin (194) divided by the total number of persons (508) or

$$\frac{(257)(194)}{508} = 98.15$$

TABLE 17.2 **Percentages by Column for the Brown (1981) Study**

	Racial Composition		
Behavior	Black	White	Mixed
Through	64	47	38
Around	36	53	62
Total	100	100	100

It is not at all unusual for the expected frequency to be a noninteger value. Normally the expected frequencies are computed to two decimal places.

The expected frequency is computed for every cell of the table. For the example, for the six cells of the table the expected frequencies are as shown in Table 17.3.

Note that the row and column margins of the table of expected frequencies are exactly the same as the observed frequencies. This mathematical necessity (within the limits of rounding error) can be used as a computational check to see whether the expected frequencies are computed correctly.

Now the observed frequency minus the expected frequency is computed for each cell. With these differences for each cell of the table the following is computed:

$$\frac{(\text{observed minus expected})^2}{\text{expected}}$$

and this quantity is added across all the cells of the table. This sum has approximately chi square distribution under the restricted model of independence. The degrees of freedom given r rows and c columns in the table are as follows:

$$\text{degrees of freedom} = (r - 1)(c - 1)$$

For the racial composition example, there are two rows and three columns. Thus, $(r-1)(c-1)$ equals 1 times 2, or 2. The chi square test of independence is

$$\chi^2[(r-1)(c-1)] = \text{sum } \frac{(\text{observed minus expected})^2}{\text{expected}}$$

The observed frequency is denoted as o and the expected frequency is denoted as e. The mathematical formula for the chi square test of independence is

$$\chi^2[(r-1)(c-1)] = \sum \frac{(o-e)^2}{e}$$

TABLE 17.3 **Expected Frequencies for the Brown (1981) Study**

	Racial Composition			
Behavior	Black	White	Mixed	Total
Through	98.15	72.34	86.51	257.00
Around	95.85	70.66	84.49	251.00
Total	194.00	143.00	171.00	508.00

The $\chi^2(2)$ for the example is 26.49. Using Appendix G, the p value for the value of χ^2 is less than .001, and so the null hypothesis that behavior and racial composition are unrelated is rejected. The differential probability of walking through racial pairs cannot be explained by chance.

For 2×2 tables (that is, a table with two rows and two columns), various measures of association were presented in Chapter 8. One such measure is the phi coefficient. As explained in Chapter 8, phi (ϕ) is a correlation coefficient. If phi is known, χ^2 can be computed directly

$$\chi^2(1) = N\phi^2$$

where N is the total number of persons in the study. So χ^2 equals the sample size times phi squared. This only applies to tests using 2×2 tables.

If the chi square is not significant, one concludes that the two variables are independent; that is, the variables are unrelated. If chi square is statistically significant, then one concludes that the variables are associated. To determine the direction of the association, one can compute percentages across rows or columns.

The fact that the degrees of freedom of the χ^2 test are $(r-1)(c-1)$ is not as mysterious as might seem. Recall that the degrees of freedom for interaction in analysis of variance take on a similar form. They equal the product of the number of levels of the first independent variable less one times the number of levels of the second variable less one. The total number of cells in the table are rc, the number of rows times the number of columns. To test for independence, the row and column margins are used. The sum of the expected frequencies must equal these row and column margins. There are r row margins and c column margins. Because both the row column margins must sum to N, there is one constraint on the row and column margins. So the number of unconstrained frequencies is the total number of cells less the number of rows and columns plus one. In terms of symbols,

$$rc - r - c + 1$$

which equals

$$(r - 1)(c - 1)$$

This equals the degrees of freedom for the χ^2 test of independence.

Assumptions

One major assumption of the χ^2 test is that observations are independent. To ensure that the assumption is met, the total N must represent that many unique responses. *The same person must not enter the table more than once.* The number of persons must equal the number of observations.

The χ^2 test of hypotheses of association between variables is only an approximate test. That is, the sum of the observed minus the expected squared

divided by the expected has only approximately a χ^2 distribution under the restricted model of independence. The p values obtained are only approximate. How good the approximation is depends, in general, on the overall sample size. The larger the sample size, the better is the approximation. A good rule of thumb is that the total N divided by the number of cells must be at least five before the approximation becomes quite good. In terms of symbols: N must be greater than or equal to $5rc$; that is, $N \geq 5rc$.

McNemar Test

The χ^2 test of independence presumes that the observations are independent. It is not at all uncommon for observations to be linked. In Chapter 13, the paired t test for scores that are linked or paired across two conditions is described. Described here is a similar procedure for linked scores in which both the independent and dependent variables are dichotomies.

Consider an election survey in which 100 persons are interviewed and 55 favor candidate A and 45 candidate B. These same 100 persons are interviewed again and asked who it is that they prefer. Now 49 prefer A and 51 prefer B. The issue is whether the percentage of those favoring the candidates has changed significantly over time. The independent variable is time, and the dependent variable is candidate preference. It would not be valid to employ a χ^2 test of independence because the same persons were interviewed in both of the surveys.

To perform the McNemar test, one examines only those who have changed over time. So, the number who switched from candidate A to B is compared with the number who switched from B to A. If the independent variable had no effect on the dependent variable, within the limits of sampling error, these two numbers should be the same. The McNemar test evaluates the null hypothesis that the two types of changers are equal. If this null hypothesis is false, the null hypothesis that the independent variable has no effect on the dependent variable also is false.

There are two key frequencies that must be determined to compute the McNemar test. The persons who switch from one category to the other category for the dependent variable must be counted. The two frequencies are designated as a and d. So, for the example, a is the number who switched from candidate A to B and d is the number who switched from B to A. The formula for the McNemar test is

$$\chi^2(1) = \frac{(|a - d| - 1.0)^2}{a + d}$$

(The expression $|a - d|$ is the absolute value of $a - d$. If $a - d$ is negative, the sign becomes positive.) The degrees of freedom for the McNemar test are one. The -1.0 term in the numerator is called the *correction for continuity*.

Such a correction improves the accuracy of the χ^2 approximation. (A similar correction was proposed for the χ^2 test of independence for 2×2 tables. Recent work has shown that the correction there is not necessary.)

Assumptions

Even though the two groups are nonindependent, all other scores must be independent. Also, the χ^2 distribution is used to approximate the sampling distribution of the McNemar test. If $a + d$ is small, the approximation is not very good. One rule of thumb is that $a + d$ must be at least ten before the test is performed. Even if the approximation were good for $a + d$ less than ten, the test would be of little use because its power would be so low.

Example

Consider the following experiment. Mita, Dermer, and Knight (1978) took one picture of 33 persons, but for each person two different pictures were developed. One was a usual or regular picture. For the other, the negative was turned upside down before printing, causing the print to represent a mirror image of the person photographed. Each person and a person's friend were asked which of the two pictures they preferred. The normal print would show the way that others see the person, and the reversed print would show how the person would see him or herself as in a mirror. According to the social psychologist Robert Zajonc, individuals generally prefer the familiar, and so friends should prefer the regular photo and the persons themselves should prefer the reversed photo.

The independent variable from this experiment is friend versus self, and the dependent variable is picture chosen, regular or reversed. Although there are 66 persons in the study, only 33 of them are independent because there are actually 33 pairs of friends. To perform the McNemar test, it must be determined how many times the friend preferred the regular picture and the self preferred the reversed picture. According to Mita and his colleagues this number should be high relative to the number of times that the friend preferred the reversed picture and the self preferred the regular picture.

The results from the experiment by Mita, Dermer, and Knight are that 15 pairs operated as predicted and 7 pairs were in the opposite direction. The McNemar test result is

$$\frac{(|15 - 7| - 1.0)^2}{15 + 7}$$

The χ^2 (1) value is 2.23. Using Appendix G, this value does not equal or exceed the value of 3.84 necessary for it to be statistically significant at the .05 level of significance. So although the results are in the predicted direction, they are not statistically significant. There is no statistically significant

evidence that persons prefer the reversed picture of self and friends prefer the normal picture.

χ^2 Goodness of Fit Test

Sometimes a researcher has a hypothesis about the distribution of a nominal variable and wishes to evaluate it. Consider the following examples:

1. In a study of extrasensory perception, a researcher asks 40 supposed psychics whether a coin that is flipped is heads or tails. Of the 40 psychics, 24 are correct and 16 are incorrect. Is this significantly better than 20 correct and 20 incorrect expected by chance?
2. A computer scientist wants to test how random her random number generator is. She has a computer generate 1000 random integers from 1 to 10. If the generator is truly random, then each integer should appear 10% of the time.
3. A researcher seeks to compare whether enough women are called for jury duty in a given county of the United States. By using census data, the researcher determines that 52% of the adult population is female. Of 458 persons called for jury duty 212 are females.

In each of these cases, there is a nominal variable. For the first, it is heads or tails; for the second, it is integer from one to ten; and for the third, it is gender. The researcher has some way of predicting the percentage of cases for each category of the nominal variable. The *expected frequency* for a category equals the total N times the proportion that is predicted for that category. So for each category of a nominal variable, there is an observed frequency and an expected frequency.

The observed frequency can be compared to the expected frequency. It turns out that the expression

$$\text{sum}\left[\frac{(\text{observed} - \text{expected})^2}{\text{expected}}\right]$$

has a χ^2 distribution with $k - 1$ degrees of freedom, where k is the number of categories of the nominal variable. If χ^2 is significant, the model or theory that predicts the distribution is incorrect in some way. If χ^2 is not significant, the frequencies are compatible with the theory.

Note that the formula for the χ^2 goodness of fit test is identical to that for the χ^2 test of independence. The difference between the two tests is in how the expected frequencies are computed.

Assumptions

The χ^2 goodness of fit test requires that observations be independent. One consequence of this assumption is that the same person may enter the table

only once. A second assumption is that all expected values must be nonzero. If theory predicts that a category has no members, a χ^2 test is not necessary. One need only see whether the category has any members. If it does the theory is falsified. For the χ^2 approximation to be adequate, expected values should be at least five.

Example

In 1866 the monk Gregor Mendel reported the results of his experiments on the inheritance of traits. Mendel took seeds that were pure strain yellow and pollinated them with pure strain green. A total of 529 plants were produced. If his theory of inheritance were correct, then 25% of the peas produced should be pure yellow, 25% pure green, and the remaining 50% should be a hybrid mixture of yellow and green.

What Mendel found was as follows:

Yellow	126
Hybrid	271
Green	132
Total	529

At issue is how well Mendel's theory of inheritance predicts the distribution of pea plant colors.

Because the theory predicts 25% yellow, 50% hybrid, and 25% green the expected frequency of plants are

$$\text{Yellow: } .25 \times 529 = 132.25$$

$$\text{Hybrid: } .50 \times 529 = 264.50$$

$$\text{Green: } .25 \times 529 = 132.25$$

Each of these expected frequencies equals the proportion predicted by the theory times the total number of cases. The sum of these expected frequencies is 529, which is what it should be.

Now these expected frequencies are compared with the observed frequencies.

Plant	*Observed*	*Expected*	*Observed–Expected*
Yellow	126	132.25	–6.25
Hybrid	271	264.50	6.50
Green	132	132.25	–.25

Note that the sum of the observed minus expected is zero, which is a mathematical necessity. So, for Mendel's data, the χ^2 is found to be

$$\chi^2(2) = \frac{(-6.25)^2}{132.25} + \frac{6.50^2}{264.50} + \frac{(-.25)^2}{132.25} = .46$$

Using Appendix G, a value of χ^2 with two degrees of freedom requires a value of 5.99 to be significant at the .05 level of significance. So $\chi^2(2) = .46$

is not statistically significant. The degrees of freedom are two because there are three categories, making k equal to three. Hence the difference between Mendel's obtained distribution of peas and the distribution expected by theory can be attributed to sampling error. The results are compatible with Mendel's theory.

Other Models for Nominal Dependent Variables

There are many more complex models for nominal dependent variables than those considered in this chapter. For instance, more than one independent variable may be present and the effect of the interaction between the two independent variables may be of interest. Or one may wish to test the effect of a three-level nominal variable on a nominal variable with nonindependent groups. To estimate and test such models, a general method called *log linear analysis* can be used (Fienberg, 1977; Reynolds, 1977).

The model for log linear analysis is formally similar to an analysis of variance model. Like the methods presented in this chapter, log linear analysis produces a set of expected frequencies which are compared to the observed frequencies. However, for most log linear models the expected frequencies require extensive computation, and therefore computers must be used. The discrepancies between observed and expected frequencies are evaluated by the χ^2 distribution. Log linear models are used primarily in survey research, but they could be applied to almost any area of research.

Summary

The methods discussed in this chapter were developed for variables measured at the nominal level of measurement, whereas the methods discussed in the previous five chapters assume that the dependent variable is measured at the interval level of measurement. These methods, as well as those for ordinal dependent variables, are called *distribution-free* methods because no assumptions are made concerning the distribution of the residual variable. There are three reasons for using distribution-free methods. First, because the dependent variable may be clearly measured at the nominal or ordinal level of measurement, the procedures developed for interval data are inappropriate. Second, the dependent variable may be at the interval level of measurement, but the researcher may be unwilling to assume that the residual variable has a normal distribution. Third, the distribution-free test evaluates a different null hypothesis from that of the distribution-tied test.

The χ^2 *test of independence* is used to evaluate association between a pair of nominally measured variables. It takes as the restricted model that there is

no association between the two variables. The χ^2 distribution is used as an approximation to evaluate the plausibility of the restricted model. It involves computing the frequencies expected given no association and comparing them with the observed frequencies. The expected frequency for a cell equals the cell's row margin times the cell's column margin divided by the total number of observations. The degrees of freedom of the test are the number of rows minus one, times the number of columns minus one.

The *McNemar test* evaluates whether a nominal independent variable affects a nominal dependent variable in which the groups are not independent. To use this test, the number of persons who switch from one category to the other is determined. The χ^2 test has one degree of freedom.

For the χ^2 *goodness of fit test* a theory predicts the relative frequencies for each category of the nominal variable. Like the χ^2 test of independence, the goodness of fit test compares observed to expected frequencies. The number of degrees of freedom is the number of categories less one.

More complicated models for nominal dependent variables can be tested through the use of *log linear models*. Like the χ^2 tests presented in this chapter, log linear models involve specifying a restricted model and making predictions concerning the expected frequencies. These expected frequencies are compared to the observed frequencies through the χ^2 distribution.

Problems

1. Locate in the χ^2 table in Appendix G the minimal value of χ^2 to achieve statistical significance.

	df	*p level*
a.	1	.05
b.	5	.01
c.	3	.001
d.	2	.05
e.	10	.01
f.	19	.10

2. For the following table compute and interpret the χ^2 test of independence.

	Male	*Female*
Yes	15	30
No	25	8
Undecided	12	20

Interpret the result.

3. A researcher seeks to compare how many women are called for jury duty in a given county of the United States, in relation to the number of men. By using census data, the researcher finds that 52% of the population

is female. Of 458 persons called for jury duty 212 are females. Are those called for jury duty representative of the general population?

4. The following table (Anderson, 1954) presents the relationship between seeing an ad and buying a product.

		See an Ad: Yes	See an Ad: No
Buy the Product	Yes	138	147
	No	118	543

Compute a χ^2 test of independence and interpret the result.

5. Below is a table of the preferences of blacks and whites to be stationed in a northern and southern camp during World War II (Stouffer, Suchman, Devinney, Star, & Williams, 1949).

		Blacks	Whites
Regional Preference	North	2027	2024
	South	2268	1717

Compute a χ^2 and interpret the result.

If one splits persons by where they were born, North versus South, one obtains the following pair of 2×2 tables.

	Area of Birth: North		Area of Birth: South	
Regional Preference	Blacks	Whites	Blacks	Whites
North	1263	1829	764	195
South	286	672	1982	1045

Compute the χ^2 test of independence separately for those born in the North and those born in the South. For each group interpret the relationship.

6. In a study of extrasensory perception, a researcher asks 40 supposed psychics whether a coin that is flipped is heads or tails. Of the 40 psychics, 24 are correct and 16 are incorrect. Is this significantly better than the 20 correct and 20 incorrect expected by chance?

7. A computer scientist wants to test how random her random number generator is. She has the computer generate 1000 random integers from 1 to 10. If the generator is truly random, then each integer should appear about 10% of the time. She finds the following results.

				Integer					
1	2	3	4	5	6	7	8	9	10
105	99	101	111	85	103	101	96	101	98

Use a χ^2 goodness of fit test to evaluate whether the ten numbers are equally likely.

8. A local politician wants to know if her popularity is improving. She had surveyed 112 persons and found that 40 thought that she was doing a good job and 72 did not. In a more recent survey, 50 thought that she was doing a good job and 62 did not. Given that the two groups are independent, test to see if her popularity is improving.

9. For problem 8, assume now that the same set of persons were interviewed at both times. The complete set of results are as follows:

Time 1	*Time 2*	*n*
good	good	35
good	poor	5
poor	good	15
poor	poor	57

Is her popularity significantly improving?

10. In problem 8 the candidate is rated as good by 40 and poor by 72 in her first survey. Test the hypothesis that as many persons like the candidate as dislike her.

11. An investigator has 27 mothers and fathers listen to recorded cries of their infant child and the cries of another child. Each parent is asked to identify the cries of their own child. The results are as follows:

Father Correct	*Mother Correct*	*n*
yes	yes	5
yes	no	1
no	yes	9
no	no	12

Are mothers better able than fathers to recognize the cries of their own infant?

12. Consider the variables of religion and support for or against abortion, where the entries represent observed frequencies.

	Religion			
Abortion Attitude	*Protestant*	*Catholic*	*Jewish*	*Other*
Approve	33	44	14	54
Disapprove	21	65	4	33

Compute a χ^2 test of independence. Interpret the results.

18 Models for Ordinal Dependent Variables

In the preceding chapter it was pointed out that statistical techniques that are used for variables measured at the interval level of measurement are not always appropriate. First, the dependent variable may clearly not be measured at the interval level of measurement. Second, the researcher may be unwilling to make the assumptions that are required for distribution-tied tests. For instance, the normality assumption may be clearly implausible. If either of these cases holds, a distribution-free test may be needed. In this chapter, the topic is the set of models for dependent variables measured at the ordinal level of measurement. As explained in Chapter 1, the ordinal level of measurement implies that the objects can only be rank ordered and that quantitative differences between pairs of objects cannot be assessed.

In this chapter all models have an ordinal dependent variable. The set of models to be considered are presented in Table 18.1. The Mann-Whitney U test is the distribution-free analog of the two-sample t test discussed in Chapter 13. The independent variable is a nominal variable with two levels. So for a Mann-Whitney test there are two groups of persons. Additionally, the dependent variable is measured at the ordinal level of measurement. The Kruskal-Wallis test is the distribution-free analog to one-way analysis of variance. There are multiple groups of persons with the Kruskal-Wallis test and so the independent variable is a nominally measured variable. Like the Mann-Whitney test, the dependent variable is measured at the ordinal level of measurement.

Both the Mann-Whitney and the Kruskal-Wallis presume that the groups are independent. If the groups are not independent, then different tests must be employed. If the independent variable is a dichotomy and the scores in each group are linked, the sign test is appropriate. The sign test's distribution-tied analog is the paired t test described in Chapter 13. If there are more than two groups that are nonindependent, the appropriate test is Friedman two-way

TABLE 18.1 **Models for Ordinal Dependent Variables**

Level of Measurement of the Independent Variable	Independent Groups	Test	Distribution-Tied Counterpart
Nominal (dichotomy)	Yes	Mann-Whitney	*t* test
Nominal (multilevel)	Yes	Kruskal-Wallis ANOVA	One-way ANOVA
Nominal (dichotomy)	No	Sign test	Paired *t* test
Nominal (multilevel)	No	Friedman two-way ANOVA	Repeated measures ANOVA
Ordinal	—	Rank-order coefficient	Correlation

ANOVA. Its distribution-tied analog is the repeated measures ANOVA, which was presented in Chapter 15.

Finally, if both the independent and dependent variable are measured at the ordinal level of measurement, then the degree of association between the two variables is measured by the *rank-order coefficient,* sometimes called *Spearman's rho*. This coefficient is the distribution-free analog to the ordinary correlation coefficient. (Because the independent variable is not nominal, it is not relevant to refer to independent or nonindependent groups.)

It is important to realize that a distribution-free method evaluates different null hypotheses than the comparable distribution-tied method. If the different groups have the same distribution but different medians, the distribution-free tests evaluate whether the groups have equal medians. If, however, the groups have different distributions, then the null hypothesis becomes more complicated to state.

For the Mann-Whitney test and Kruskal-Wallis ANOVA, the general null hypothesis is that the groups, when considered as a single sample, all have mean percentile ranks of 50.0. For the sign test and Friedman two-way ANOVA, the null hypothesis is that, for each pair of conditions, persons are just as likely to have a higher score in one condition as they are in the other. When presenting these tests, for reasons of simplicity the null hypothesis will be stated that the groups have equal medians. It should be remembered that when the distributions are different, the null hypothesis is more complicated.

The rank-order coefficient measures any consistent positive or negative

relationship between a pair of variables. The ordinary correlation coefficient, or r, measures only the linear association between a pair of variables. So, the null hypothesis for Spearman's rho is no positive or negative relationship between the variables.

The procedures that are to be presented in this chapter for ordinal data presume that there are no ties. If there are ties, then each score is given the mean of the tied rank. So, the following set of scores

5, 6, 6, 6, 9, 9, 12, 13, 13, 13, 13

would yield ranks of

1, 3, 3, 3, 5.5, 5.5, 7, 9.5, 9.5, 9.5, 9.5

Methods that correct for ties in the ranks for formulas described in this chapter are described in more advanced texts (Bradley, 1968; Siegel, 1956). However, not correcting for ties when there are not many seems to have little effect on the p values.

When working with ranks there are two useful computational checks. The first is to make sure that the last rank (given that it is not tied) equals n, the sample size. If it does not, there is an error in the ranking. The second computational check is to compute the mean of the ranks. It should equal $(n + 1)/2$, even if there are tied ranks. If the mean of the ranks does not equal $(n + 1)/2$, there is an error in assigning ranks.

Mann-Whitney U *Test*

The Mann-Whitney U test is analogous to the two-sample t test described in Chapter 13. However, the assumptions concerning normal distribution and homogeneity of variance are not made by the Mann-Whitney U test. It is then a distribution-free "t test." The test is fairly commonly used in medicine and the biological sciences, but it is relatively infrequently used by most social scientists. Nonetheless, it is an appropriate test when the assumption of normality seems totally implausible.

The Mann-Whitney U test evaluates not the similarity of the means of two groups but rather any consistent difference in the mean percentile scores of the two groups. If the two groups have similar distributions, the Mann-Whitney evaluates whether the two groups have equal medians. Because the mean and median may not be the same, even in the population, the t test and the Mann-Whitney test do not evaluate exactly the same null hypothesis.

The Mann-Whitney test begins with a ranking of all the scores ignoring the fact that the persons are in two different groups. Persons are therefore treated as if they were members of one large group. The ranks then are averaged for the persons in each of the groups, and the difference is computed. This difference between the ranks in the two groups will be denoted as Q. At issue

is whether the difference between ranks is much larger than it would be if the ranks were assigned randomly.

One way to determine the unusualness of the value of Q is through random assignment of a rank to each person. That is, the actual data are ignored and the subjects are rank ordered again, but this time the ranking is done randomly. Then with these random ranks, the value of Q is computed. If this were done repeatedly, one would obtain a distribution for Q. One would then determine just how unusual the obtained value of Q is relative to the values obtained for Q by using a random procedure. This is the essence of the Mann-Whitney U test. It essentially computes the difference between the average rank for the persons in the two groups and judges whether that difference between ranks could have occurred by chance. It does this by comparing the obtained value of Q to what the value of Q would be if persons were randomly assigned ranks.

Consider the following simple case: The number of persons in each group equals three. The data for the two groups are

Group 1: 12, 19, 18

Group 2: 25, 23, 30

The six scores are rank ordered from smallest to largest, as follows:

Group 1: 1, 3, 2

Group 2: 5, 4, 6

The mean or average rank is 2.0 for group 1 and 5.0 for group 2. The difference between the mean rank of group 1 from the mean rank of group 2 is 3.0. At issue is how unusual the value of 3.0 is. If ranks were randomly assigned to each of the six persons, the mean rank difference could be computed. If done enough times, the following mean rank differences with the following probabilities would be obtained.

Difference in Mean Rank	*Probability*	*Cumulative Probability*
3.00	.05	.05
2.33	.05	.10
1.67	.10	.20
1.00	.15	.35
.33	.15	.50
– .33	.15	.65
–1.00	.15	.80
–1.67	.10	.90
–2.33	.05	.95
–3.00	.05	1.00

For instance, a difference between mean ranks of 1.00 or greater for two groups of size three would occur by chance 35% of the time. It can be seen that the obtained value of the difference between ranks of 3.0 is unusual and

would occur by chance only 5% of the time. Because a value of –3.0 would also occur by chance 5%, the two-tailed p value is .10.

For the Mann-Whitney test, the average difference between the ranks is not the statistic that is computed but rather a statistic that could be used to derive it. The statistic computed is the sum of the ranks of the group with the smaller sample size. To see that the sum of the ranks of one group yields the difference between mean ranks, consider the example of two groups of size three. If the sum of the ranks for one group is R, then it is a mathematical necessity that the mean difference between ranks must be $2R/3 - 7$. So if R is six, the mean rank difference must be –3.0. The advantage of the sum of the ranks over the mean rank difference is that the sum of the ranks is always a positive integer, whereas this is not true of the mean rank difference. This fact makes it much easier to table the sum of the ranks rather than the mean difference.

To conduct a Mann-Whitney one proceeds as follows. All of the numbers are rank ordered from smallest to largest. Then the ranks in the smaller sized group are summed (not averaged). The sum of the ranks in the smaller sized group is denoted as R. If both groups have the same sample size, the sum of ranks of either group can be used. The sample size of the smaller group is denoted as n_1 and the sample size of the larger group as n_2. The Mann-Whitney test statistic, U is:

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2.0} - R$$

where R is the sum of the n_1 ranks. The value of U ranges from zero to $n_1 n_2$. If the ranks are, on average, larger in the n_1 group, U is small. If the ranks are larger, on average, in the n_2 group, then U is large. So if the restricted model of equal medians is false, the value of U is either very large or very small. To determine whether U is unusually large or small depends on the sample sizes. If both n_1 and n_2 are less than or equal to 20, tables are used. If either is greater than 20, an approximation is used.

Both n_1 and n_2 Are Less than or Equal to 20

First the value of U is computed. Then the obtained value is compared to those values tabled in Appendix H. If the value of U is greater than or equal to the larger value in the table or smaller than or equal to the smaller value in the table, then the restricted model that the medians of the two groups are equal is rejected. So for instance, if $n_1 = n_2 = 10$, the value of U must exceed or equal 78 to be significant at the .05 level or be less than or equal to 28. In Appendix H, the smaller sample size n_1 is the first column, and n_2 is the second column. Four significance levels are given: .10, .05, .02, and .01.

Either n_1 or n_2 Is Greater than 20

In this case, one does not use the tables in Appendix H, but rather relies on the fact that as the sample sizes increase, the distribution of U approaches the normal distribution, with a mean of

$$\frac{n_1 n_2}{2}$$

and a standard deviation of

$$\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

Using these facts, U is converted into a variable that has approximately a standard normal distribution under the restricted model. This quantity is denoted as Z_U. That is, from U its theoretical mean is subtracted and the difference is divided by its theoretical standard deviation. The complete formula is

$$Z_U = \frac{U - n_1 n_2/2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}}$$

Although the formula looks complicated, it involves only the sum of the ranks and n_1 and n_2. The quantity Z_U has a standard normal or Z distribution. That is, given the restricted model, the statistic is approximately normally distributed, with a mean of zero and variance of one. Appendix C can be used to determine the p value. The value closest to Z_U (ignoring sign and rounding down) is located. Then take the probability for Z and subtract it from .5, and multiply this difference by two. So for $Z_U = -2.51$, the probability is .4838. The p value is $(.5000 - .4838) \times 2 = .0324$.

As was stated earlier the statistic is only approximately normally distributed. This means that the p values are only approximate. How good the approximation is depends on n_1 and n_2. As they get larger, the approximation gets better.

Examples

Consider the data in Table 18.2. Because the groups have the same sample size, either group's ranks can be summed. The sum of the ranks in group A is 71. Because n_1 and n_2 are both seven, the value of U is

$$(7)(7) + \frac{7(7 + 1)}{2} - 71 = 6$$

Looking this value up in Appendix H, a value of U of six with n_1 and n_2

TABLE 18.2 **Example of Mann-Whitney Test**

Group A		Group B	
Score	Rank	Score	Rank
23	5	34	8
43	9	19	4
53	11	11	1
64	13	13	2
27	7	25	6
82	14	18	3
63	12	51	10

equal to seven is statistically significant at the .05 level. This value indicates that the groups' distributions are significantly different.

Assume that there are two groups, $n_1 = 18$, $n_2 = 22$, and $U = 246$. Given $n_1 = 18$ and $n_2 = 22$, the expected mean is

$$\frac{(18)(22)}{2} = 198$$

The variance is

$$\frac{(18)(22)(18 + 22 + 1)}{12} = 1353$$

The square root of 1353 is 36.78. The test that U does not differ from its population mean is

$$Z = \frac{246 - 198}{36.78} = 1.31$$

which is not significant. Therefore the distributions of the two groups do not significantly differ.

Sign Test

Although the Mann-Whitney test does not presume normality or homogeneity of variance, it is still required that the scores be independent from one another. It may happen that scores are paired, as described for the paired t test in Chapter 13. Each score in a given group is paired or linked to one and only one score in the other group. Scores can be paired because they come from the same person, come from a couple such as friends or littermates, or come from two persons who interact with each other.

A procedure called the *sign test* can be used to test hypotheses about the medians of two samples whose scores are linked. The sign test is very simple.

The two conditions are denoted as I and II, and it will be assumed that the same person is in both conditions. (As with any nonindependent group design, it need not be person that links together the pair of scores, but persons are used in the illustration.) If a person has exactly the same score in condition I as condition II, that person's scores are dropped from the analysis and the n is reduced by one. Like the paired t test, a difference is computed for each person. So the condition I score is subtracted from condition II score. The number of scores with positive *signs* is denoted as c. If n, the number of untied cases, is less than or equal to 25, then Appendix I is used to determine significance.

If n is greater than 25, the following Z approximation is used.

$$Z = \frac{|2c - n| - 1.0}{\sqrt{n}}$$

where n is the number of persons who have different scores and c is the number of persons whose difference score is positive. (The expression $|2c - n|$ is the absolute value, so the sign of $2c - n$ is always positive.) To determine the p value, the probabilities in Appendix C are used. (See the discussion of Z_U in the Mann-Whitney section.)

As an example, each of ten nine-year-old children work with a seven-year-old child on a task. Observers rate the degree of creativity for each child on the task. The hypothesis is that nine-year-olds are more creative than seven-year-olds. The data are as follows:

Pair	*Nine-Year-Old*	*Seven-Year-Old*
1	7	5
2	6	4
3	5	5
4	8	4
5	7	3
6	8	5
7	4	6
8	7	3
9	9	6
10	6	4

First, it is noted that because the two scores are the same for pair 3, that pair is dropped from the analysis. The n now becomes nine. The difference scores are 2, 2, 4, 4, 3, –2, 4, 3, and 2. Of these nine differences, eight are positive. So n is nine and c is eight. Using Appendix I for these values, the result is statistically significant at the .05 level. So, it is concluded that the nine-year-olds are more creative than the seven-year-olds.

As a second example consider 45 persons who entered a smoking reduction program. One year later, 27 persons have reduced their amount of smoking but 18 increased. Because n is greater than 25, the Z method is used. The value of Z is

$$Z = \frac{|(2)(27) - 45| - 1.0}{\sqrt{45}}$$

which equals 1.19 with a p value of .234, which is not statistically significant at the .05 level of significance. Thus the number of persons that reduced their smoking is not significantly greater than the number that increased.

Kruskal-Wallis Analysis of Variance

The Mann-Whitney test is limited to a dichotomous independent variable, whereas the Kruskal-Wallis test allows for multilevel independent variables. Its distribution-tied cousin is one-way analysis of variance. Although Kruskal-Wallis and Mann-Whitney appear to be very different, it is a statistical fact that Mann-Whitney is a special case of Kruskal-Wallis.

Like the Mann-Whitney U test, the Kruskal-Wallis test begins with a ranking of all of the data from smallest to largest. The ranks are summed in each group. It is the sum of these ranks that are analyzed. Also, like the Mann-Whitney test, the Kruskal-Wallis test evaluates whether the groups have any consistent differences in mean percentile rank.

The formula for the Kruskal-Wallis analysis of variance, called H, is

$$H = \left[\frac{12}{N(N+1)}\right]\left[\sum \frac{R_j^2}{n_j}\right] - 3(N+1)$$

where N is the number of persons across groups, k is the number of groups, n_j is the sample size in the jth group, and R_j is the sum of the ranks in the jth group. The quantity H is approximately distributed as χ^2 with $k-1$ degrees of freedom under the restricted model that the medians of all the groups are equal. Hence a significant χ^2 indicates that the groups differ in their medians. It has been found that this χ^2 approximation is quite good if the sample size in every group is at least five.

The formula for the Kruskal-Wallis test looks bewildering. Actually its rationale, if not its derivation, is quite simple. Imagine that the scores are first ranked. Then using the ranks, a one-way ANOVA is computed. From this one-way ANOVA the mean squares for groups would be computed. Such a mean square would take the total of the ranks and square it. These terms are present in the Kruskal-Wallis formula. Also, the $3(N+1)$ term in the formula is analogous to the correction term for the mean in ANOVA. This mean square for groups is not divided by the mean square for persons within groups but rather by a population variance for groups, and that is why the distribution is χ^2 and not F (see Chapter 11). The population variance can be determined because the scores are ranks, and the variance is therefore known.

Consider the data in Table 18.3. The sums of the ranks in the three groups are 63, 30, and 78. The Kruskal-Wallis statistic is

TABLE 18.3 **Kruskal-Wallis Analysis of Variance**

Group A	Rank	Group B	Rank	Group C	Rank
24	4	19	2	53	17
29	7	21	3	46	14
34	10	36	11	39	12
47	15	17	1	42	13
31	9	30	8	50	16
55	18	25	5	28	6
Sum	63		30		78

$$H = \left[\frac{12}{18(18 + 1)}\right]\left[\frac{63^2}{6} + \frac{30^2}{6} + \frac{78^2}{6}\right] - 3(19) = 7.05$$

Using the χ^2 distribution with two degrees of freedom, the value of 7.05 is statistically significant at the .05 level of significance. So, the restricted model that the groups have the same population medians is rejected.

Friedman Two-Way ANOVA

As described in Chapter 15, a design in which each person is at each level of a nominal independent variable is called repeated measures ANOVA. Here, the distribution-free analog to repeated measures ANOVA is presented. It is called the *Friedman two-way ANOVA*.

As in repeated measures ANOVA, each person is at every level of the independent variable or observations are linked across conditions in some way. However, with the Friedman test, the null hypothesis is that the groups' medians, as opposed to the groups' means, are equal.

To conduct a Friedman ANOVA, scores are separately ranked for each person. This is different from the Kruskal-Wallis, where the entire set of scores is ranked. The formula for the Friedman test is

$$\left[\frac{12}{nk(k + 1)}\right]\sum R_j^2 - 3n(k + 1)$$

where n is the number of persons in the study, k is the number of conditions, and R_j is the sum of the ranks for condition j.

When there are two conditions and k is two, the Friedman test is essentially identical to the sign test. Two minor adjustments should be made. First, the square root of the Friedman χ^2 should be taken to make it comparable to the sign test Z value. Second, the square root of χ^2 is slightly larger than the Z value of the sign test because the formula for sign test Z has a 1.0 value

subtracted, which is not the case in the Friedman test. If there are just two groups, the sign test should be preferred.

The data in Table 18.4 were presented earlier in this chapter. However, in this instance it is assumed that the data are from six subjects, each of whom is in every condition.

In the table, the three conditions are rank ordered for each subject. The sum of the ranks in the three conditions are 13, 7, and 16. The n is six and k is three. The Friedman statistic is

$$\left[\frac{12}{18(3+1)}\right](13^2 + 7^2 + 16^2) - 3(24) = 7.00$$

Using the χ^2 distribution with two degrees of freedom, this value is statistically significant at the .05 level of significance.

Spearman's Rank-Order Coefficient

In Chapter 8 Spearman's rank-order coefficient was described. It is a measure of association between two ordinally measured variables. Spearman's rank-order coefficient is denoted by r_S. Its formula is the standard correlation coefficient applied to ranks. For this measure the scores for each variable are separately rank ordered. Like r, r_S can vary between -1 and $+1$, and zero indicates that there is no association between the two variables.

There are three major reasons for employing the rank-order coefficient instead of the distribution-tied test. First, the data may be truly ordinal, and not interval as assumed by the ordinary correlation coefficient. Second, it is useful in cases where it cannot be assumed that the variables have a normal distribution. Third, the relationship between the two variables may not be exactly linear. If as X increases, Y increases but in a nonlinear fashion, then the rank-order coefficient may be a more appropriate measure of association than the ordinary correlation coefficient.

TABLE 18.4 **Friedman Two-Way ANOVA**

Person	Group A	Rank	Group B	Rank	Group C	Rank
1	24	2	19	1	53	3
2	29	2	21	1	46	3
3	34	1	36	2	39	3
4	47	3	17	1	42	2
5	31	2	30	1	50	3
6	55	3	25	1	28	2
Sum		13		7		16

As discussed in Chapter 8, the rank-order coefficient is an actual correlation between ranks. Besides actually correlating the ranks, there is a computationally simpler formula for the rank-order coefficient. It is based on the difference between each pair of ranks for all persons. The formula is

$$r_S = 1 - \frac{6 \sum D_i^2}{n(n^2 - 1)}$$

where n is the sample size and D_i is the difference between ranks for person i. This formula presumes that there are no ties. If there are ties, the ranks should be correlated using the regular formula for a correlation.

To evaluate whether the rank-order coefficient is significantly different from zero, the distribution of r_S under the restricted model of no association can be obtained by randomly assigning the ranks to one of the two variables. Consider the following pairs of scores.

Person	*X*	*Y*
1	5	3
2	8	9
3	6	4
4	4	1

If the scores are ranked separately for each variable, the following set of ranks would be obtained.

	Rank of	
Person	*X*	*Y*
1	2	2
2	4	4
3	3	3
4	1	1

Thus, there is perfect correspondence in the ranks, and the rank order coefficient is 1.0.

To determine how unlikely a value of 1.0 is, the sampling distribution of r_S for $n = 4$ is derived. The complete set of possible ranks is enumerated; there are a total 24 possible ranks. These 24 are listed by column, as follows:

1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	3	4	4	4	4	4	4
2	2	3	3	4	4	1	1	3	3	4	4	1	1	2	2	4	4	1	1	2	2	3	3
3	4	2	4	2	3	3	4	1	4	1	3	2	4	1	4	1	2	2	3	1	3	1	2
4	3	4	2	3	2	4	3	4	1	3	1	4	2	4	1	2	1	3	2	3	1	2	1

Using these ranks for the X variable and the ranks 1, 2, 3, and 4 for the Y variable, these 24 pairs of ranks produce all the possible rank-order coefficients. There are eleven *different* rank-order coefficients with the following frequencies:

Rank-Order Coefficient	*Frequency*	*Probability*
1.0	1	.04125
.8	3	.12375
.6	1	.04125
.4	4	.16500
.2	2	.08250
.0	2	.08250
−.2	2	.08250
−.4	4	.16500
−.6	1	.04125
−.8	3	.12375
−1.0	1	.04125

So, the obtained rank-order coefficient of 1.0 would occur by chance only 4.125% of the time. Allowing for a perfect negative rank-order coefficient, the exact p value is .0825.

Fortunately, it is not necessary to do all this work. Tables and approximations are used to test r_S. The procedure used to test whether r_S is equal to zero in the population depends on the sample size. If n is less than or equal to 30, one uses the table in Appendix J. If the observed value of r_S equals or exceeds the tabled value in Appendix J, the value of r_S is statistically significantly different from zero at the appropriate level of significance. So for example, if n is 15 and r_S is .31, it does not exceed the critical values in Appendix J, and so the correlation is judged not to be statistically significant.

If n is greater than 30, one uses the ordinary test of a correlation coefficient.

$$t(n\text{–}2) = \frac{r_S\sqrt{n-2}}{\sqrt{1-r_S^2}}$$

where r_S is the Spearman rank-order coefficient and n the sample size. So, if n is greater than 30, the t distribution is used as the test statistic. The use of the formula is only an approximation. How good the approximation is depends on n. As n gets larger, the approximation gets better.

To illustrate the computations, consider the following example. A total of twelve countries are rank-ordered on their economic wealth and their rate of literacy. The results are

Country:	A	B	C	D	E	F	G	H	I	J	K	L
Wealth:	5	8	10	12	1	9	3	6	4	2	7	11
Literacy:	6	8	11	12	1	7	3	4	5	2	9	10

The sum of the discrepancies squared is 16 and the rank-order coefficient is

$$1 - \frac{6(16)}{12(12^2-1)} = .944$$

Using the table in Appendix J, it is found that a .944 coefficient with an n of 12 is statistically significant at the .002 level. Thus, the association between wealth and literacy cannot be explained by chance.

If 40 persons' intelligence and athletic ability are ranked, the rank-order coefficient might be .125. The test of this correlation is

$$t(38) = \frac{.125\sqrt{40-2}}{\sqrt{1-.125^2}} = .777$$

This value is not statistically significant at the .05 level.

Power Efficiency

To measure the relative power of a distribution-free and a distribution-tied method, statisticians have developed a measure called *power efficiency*. Assume that there are two tests A and B and A is the more powerful test. Let n_a be the number of subjects needed to achieve a given level of power for test A and n_b be the number of subjects needed for test B to achieve the same power as test A with n_a observations. Because test A is more powerful than test B, n_a must be less than n_b. The power efficiency of test B in relation to test A is defined as

$$100 \times \frac{n_a}{n_b} \text{ percent}$$

So, if the power efficiency of a given distribution-free test is 50%, one would need twice as many subjects for the distribution-free test to achieve the same power as with the distribution-tied test.

The power of the Mann-Whitney is comparable to the power of the standard two-sample t test. When the set of assumptions hold for the two-sample t test, the power efficiency of the Mann-Whitney test for moderate samples is about 95%. This value indicates that there is little loss of power in employing the Mann-Whitney U test instead of the t test when the distribution is normal. There exist certain types of distributions for which the Mann-Whitney U test has a power efficiency greater than 100%.

The power efficiency of the sign test compared to the paired t test depends on the sample size. For very small sample sizes ($n = 6$), the power efficiency of the sign test is 95.5%. For very large samples, the power efficiency drops to 63.7%.

The power of the Kruskal-Wallis test is measured in its efficiency versus the F test from an analysis of variance. When the set of assumptions hold for the F test, the power efficiency of the Kruskal-Wallis test is about 95%. There is little loss of power in employing the Kruskal-Wallis test even when the assumptions required by analysis of variance apply. This 95% figure refers to

the normal distribution. There exist certain types of distributions for which the Kruskal-Wallis test has a power efficiency of greater than 100%.

The power efficiency of the Friedman two-way ANOVA depends on the number of conditions and the number of subjects. If there are only two conditions and many subjects, the power efficiency of the test can be as low as 63.7%. If there are either very few subjects or very many conditions, the power efficiency of the Friedman test can be as high as 95.5%.

The power efficiency of the rank-order coefficient is 91% compared to the ordinary correlation coefficient. So when the assumptions necessary for computing the ordinary correlation coefficient are true and r is computed, one needs 91% of the subjects to have the same power to be able to reject the null hypothesis as one would need if Spearman's rho were computed. When the assumptions necessary for r do not hold, the power efficiency of the rank-order coefficient may be almost as good as that of the Pearson r, and in some cases it is even better.

Summary

When the dependent variable is a set of ranks or when one is unwilling to make the assumptions required in distribution-tied statistics, tests that require only variables at the ordinal level of measurement are useful.

The *Mann-Whitney test* is used to test whether a two-level independent variable affects an ordinally measured dependent variable. The test primarily evaluates whether the two groups have the same median. All the scores are ranked and the average rank of the two groups is compared. If the number of observations in both groups is less than or equal to 20, a table is used to determine statistical significance. If not, an approximation to the standard normal distribution is used.

When the independent variable is a dichotomy and the dependent variable is set of ranks and the two groups are nonindependent, the *sign test* is appropriate. The sign test involves determining which observation is larger. If the number of paired observations is less than or equal to 25, a table is used; and if greater than 25, a χ^2 approximation is used.

The *Kruskal-Wallis test* is an extension of the Mann-Whitney test when there are more than two groups. Like the Mann-Whitney test, all the scores are initially ranked, and then analyzed. The test statistic, called H, is evaluated by a χ^2 approximation. The degrees of freedom for the test are the number of groups less one.

When there are multiple groups that are nonindependent, *Friedman two-way ANOVA* can be employed. This test involves a ranking of the scores separately for each subject and then using a χ^2 approximation. Its distribution-tied analog is a repeated measures ANOVA.

The *rank-order coefficient* is used to measure association between two ordinally measured variables. The scores for each variable are first rank

ordered. The rank-order coefficient is a standard correlation of these ranks. With this measure the relationship between the variables need not be exactly linear. If the sample size is less than or equal to 30, a table is used to determine statistical significance. If n is greater than 30, the standard t test of a correlation can be used to approximate the p value.

Distribution-free tests make weaker assumptions about the data. They do have somewhat less power than distribution-tied methods when the assumptions made by distribution-tied methods are true. However, the power efficiency of distribution-free tests is usually in the mid-90s. That is, the comparable distribution-tied test has the same power as the distribution-free test with about 95% of the subjects.

Problems

1. For the following data compute a rank-order coefficient, test it, and interpret the results.

Person:	1	2	3	4	5	6	7	8
X:	7	9	11	3	12	4	5	16
Y:	10	7	6	12	4	5	8	3

2. Perform a Mann-Whitney U test for the following data set

   A: 15, 21, 28, 17, 31, 24, 18

   B: 19, 7, 15, 8, 12, 19, 10

3. For the following data compute a rank-order coefficient, test it and interpret the results.

Person:	1	2	3	4	5	6
X:	10	4	10	3	15	5
Y:	6	7	6	11	3	8

4. Using a Kruskal-Wallis analysis of variance, test whether the groups' medians differ.

   I: 11, 18, 19, 24, 31

   II: 19, 27, 15, 8, 13

   III: 15, 12, 21, 29, 17

5. An experimenter investigated the success of three methods of lowering the level of cholesterol in the blood. Using a Kruskal-Wallis analysis of variance, test whether the groups' medians differ.

   I: 111, 128, 190, 214, 198

   II: 193, 207, 125, 88, 103, 176

   III: 150, 152, 221, 129, 171

6. Twelve subjects were measured before and after psychotherapy on an adjustment scale. Higher scores indicate greater adjustment. The numbers are as follows:

Subject	*Before*	*After*
1	23	32
2	27	25
3	31	40
4	32	31
5	26	38
6	25	29
7	25	31
8	24	24
9	33	40
10	22	34
11	36	38
12	29	25

Using a distribution-free test, evaluate whether persons improved after psychotherapy.

7. Subjects were asked to lift three weights and rank order them from lightest to heaviest. All three weights were identical in objective weight, but they differed in shape: spherical, conical, and cubical. For the 20 subjects the results were as follows.

	Spherical	*Conical*	*Cubical*
Heaviest	3	5	12
Middle	9	4	7
Lightest	8	11	1

The numbers in the table indicate the number of subjects who gave the object that rank. For example, 11 subjects felt that the conical object was the lightest. Do the three objects differ significantly in perceived weight?

8. The following scores are taken from a study that compared two different methods of increasing vocabulary. The scores of ten persons, five under each method, are

   A: 16, 19, 20, 18, 24

   B: 12, 15, 16, 15, 14

   On the basis of a distribution-free test, is there any evidence that one method is superior to the other?

9. A program is developed to improve the intelligence (IQ), scores of preschool children. Two groups of children are randomly formed. Using a distribution-free test, test whether the program affects IQ score.

   Treated group: 109, 123, 141, 119, 133, 117, 118, 120

   Control group: 106, 103, 114, 120, 116, 107, 98

10. Twenty persons are randomly assigned to one of two treatments. In the treatment group, ten persons are taught a series of strategies to improve their memory. The control group learned none of the strategies. The scores on a memory test are

    Memory group: 88, 76, 83, 75, 64, 80, 76, 73, 84, 78

    Control group: 84, 73, 84, 78, 68, 78, 71, 70, 80, 79

    Using a distribution-free test, are the two groups different?

11. A psychologist studies the degree of happiness of people at various stages in life. His measure of general happiness varies from 0 to 60. In one study he compared the happiness of married and single men aged 25. Using a distribution-free test, is there a significant difference between the two groups?

Married	*Single*
58	57
45	44
50	59
54	44
49	39
39	60
50	44
51	

12. Nine persons were asked to rate the taste of cola A and cola B on a scale from one to ten. Using a distribution-free test, do persons significantly prefer one drink to the other?

Person	*Cola A*	*Cola B*
1	7	7
2	8	9
3	8	7
4	9	5
5	10	9
6	9	7
7	8	6
8	8	10
9	7	8

13. A psychologist is interested in the relationship between handedness and athletic ability. He measures the athletic ability of three groups of persons: left-handed, right-handed, and ambidextrous. His results are:

    Left-handed: 11, 13, 14, 13, 15

    Right-handed: 10, 8, 7, 10, 14

    Ambidextrous: 12, 8, 6, 11, 15

    Do a Kruskal-Wallis ANOVA to determine whether the groups significantly differ.

14. Problem 7 in Chapter 14 described a study of the effectiveness of three different treatments in relieving headache pain. The drugs studied were aspirin, acetaminophen, and a placebo. Ten different persons took one drug and rated their pain on a ten-point scale after three hours. The scores were

    Aspirin: 7, 6, 9, 5, 3, 5, 3, 2, 4, 2

    Acetaminophen: 5, 8, 6, 4, 7, 4, 6, 2, 3, 7

    Placebo: 9, 7, 8, 7, 5, 4, 6, 8, 3, 7

    Using a distribution-free test, evaluate whether the groups significantly differ.

15. A researcher seeks to compare the marital satisfaction of women who have been married for varying number of years. She finds the following (higher numbers, greater satisfaction).

    One Year: 56, 48, 57, 41

    Two Years: 63, 51, 65, 54

    Ten Years: 70, 61, 55, 58

    Using a distribution-free test, evaluate the effect of length of marriage on satisfaction.

16. The following data are taken from Diehl, Kluender, and Parker (1985).

Subject	*I*	*II*	*III*
DS	20	19	21
MM	21	18	20
JH	28	24	31
JS	17	6	10
MC	15	13	10
CM	20	13	18
LG	28	22	22
LD	30	24	26
RL	23	18	17
TW	29	19	21
TA	30	16	25
VS	34	29	32
CJ	21	20	20

    Using a distribution-free test, test for an effect due to condition.

17. In a study involving 20 experimentals and 25 controls:

    a. The sum of the ranks of the 20 experimentals is 248. Do a Mann-Whitney test to determine if the groups' distributions differ.
    b. What would be your answer if the sum of the ranks was 342?

18. For the following values of the rank-order coefficient and n, state whether the correlation is significantly different from zero.

	r_S	n
a.	−.21	78
b.	−.45	42
c.	.71	12
d.	.35	20
e.	.17	99
f.	.47	29
g.	.46	33
h.	−.19	17

Postscript

In Chapter 1 we began our journey. We have traveled through a sea of numbers, terms, formulas, and tables. Research in the social and behavioral sciences brings with it a bewildering array of symbols and terminology. If used properly, they can help us understand why human beings are the way they are. But even more important, they can help us understand how we can come to be more than what we are today.