Analysis of Categorical Data

10

Chapter Outline
10.1 The Goodness-of-Fit Test
10.2 The 2 by 2 Contingency Table
10.3 The *r* by *c* Contingency Table
10.4 Multiple 2 by 2 Contingency Tables

In this chapter, we present some additional nonparametric tests that are used with nominal, ordinal, and continuous data that have been grouped into categories. The data in this chapter are presented in the form of frequency or contingency tables. In Chapter 3, we demonstrated how one- and two-way frequency tables could be used in data description. In this chapter, we show how frequency or contingency tables can be used in the test of whether or not the distribution of the variable of interest agrees with some hypothesized distribution or whether or not there is an association among two or more variables. For example, in the material on the normal distribution in Chapter 5, we examined the distribution of blood pressure. In this chapter, we show how to test the null hypothesis that the data follow a particular distribution. In Chapter 4, we considered the association between birth weight and the timing of the initiation of prenatal care. In this chapter, we show how to test the null hypothesis that an association exists between two discrete variables versus the alternative hypothesis that there is no association. A goodness-of-fit statistic is used to test these hypotheses, and it follows a chi-square distribution if the null hypothesis is true.

10.1 The Goodness-of-Fit Test

The *goodness-of-fit* test can be used to examine the fit of a one-way frequency distribution for X, the variable of interest, to the distribution expected under the null hypothesis. This test, developed in 1900, is another contribution of Karl Pearson, also known for the Pearson correlation coefficient. The X variable is usually a discrete variable, but it could also be a continuous variable that has been grouped into categories.

To facilitate the presentation, we shall use the following symbols. Let O_i represent the number of sample observations at level i of X and E_i represent the expected number of observations at level i, assuming that the null hypothesis is true. The E_i are found by multiplying the population probability of level i, π_i, by n, the total number of observations. Since the sum of the π_i is one, the sum of the E_i is n.

A natural statistic for this comparison would seem to be the sum of the differences of O_i and E_i — that is, $\Sigma(O_i - E_i)$. However, since both the O_i and the E_i sum to n, the

sum of their differences is always zero. Thus, this statistic is not very useful. However, the sum of the squares of the differences, $\Sigma(O_i - E_i)^2$, will be different from zero except when there is a perfect fit. Squaring the differences is the same strategy used in defining the variance in Chapter 3.

One problem remains with $\Sigma(O_i - E_i)^2$. If the sample size is large, even very small differences in the observed and expected proportions at each level of X become large in terms of the O_i and E_i. Therefore, we must take the magnitude of the O_i and E_i into account. Pearson suggested dividing each squared difference by the expected number for that category and using the result, $\Sigma(O_i - E_i)^2/E_i$ as the test statistic. It turns out that this statistic, for reasonably large values of E_i, follows the chi-square distribution. In the early 1920s, Sir Ronald A. Fisher showed that this statistic has $k - 1 - m$ degrees of freedom, where k is the number of levels of the X variable and m is the number of estimated parameters. For the chi-square distribution to apply, no cell should have an expected count that is less than five times the proportion of cells with E_i that are less than five (Yamold 1970). For example, if k is 10 and two cells have expected counts less than five, then no expected cell count should be less than one (= 5[2/10]). If some of the E_i are less than this minimum value, categories with small expected values may be combined with adjacent categories. The combinations of categories must make sense substantively; otherwise the categories should not be combined.

Note that the goodness-of-fit test is a one-sided test. Only large values of the chi-square test statistic will cause us to reject the null hypothesis of good agreement between the observed and expected counts in favor of the alternative hypothesis that the observed counts do not provide a good fit to the expected counts. Small values of the test statistic support the null hypothesis.

We consider the following two examples: In the first example, no parameter estimation is required, and two parameters are estimated in the second example.

Example 10.1

(No Parameter Estimation Required): The study of genetics has led to the discovery and understanding of the role of heredity in many diseases — for example, in hemophilia, color-blindness, Tay-Sachs disease, phenylketonuria, and diabetes insipidus (Snyder 1970). The father of genetics, Abbe Gregor Mendel, presented his research on garden peas in 1865, but the importance of his results was not appreciated until 1900. One of Mendel's discoveries was the 1 : 2 : 1 ratio for the number of dominant, heterozygous, and recessive offspring from hybrid parents — that is, from parents with one dominant and one recessive gene.

Although doubts have been raised about Mendel's data, we shall use data from one of his many experiments. Table 10.1 shows the number of offspring by type from

Table 10.1 Mendel's data on garden peas: number of smooth and wrinkled offspring from hybrid parents.

AA	Aa	aa	Total
138	256	126	529

the crossbreeding of smooth seeds, (*A*), the dominant type, with wrinkled seeds, (*a*), the recessive type (Bishop, Fienberg, and Holland 1975). Dominant means that when there are both a smooth and a wrinkled gene present, the pea will be smooth. The pea will be wrinkled only when both genes are wrinkled.

The question of interest is whether or not this experiment supports Mendel's theoretical ratio of 1 : 2 : 1. The null hypothesis is that the observed data are consistent with Mendel's theory. The alternative hypothesis is that the data are not consistent with his theory. Let us test this hypothesis at the 0.10 significance level.

We must first calculate the expected cell counts for this one-way contingency table. Since the expected counts are based on the theoretical 1 : 2 : 1 ratio, the ratio tells us that we expect 1/4 of the observations to be *AA*, 2/4 to be *Aa* or *aA*, and 1/4 to be *aa*. One-fourth of 529 is 132.25, and one-half of 529 is 264.5; therefore, the expected counts are 132.25, 264.5, and 132.25, respectively. The test statistic is

$$X^2 = \frac{(138-132.25)^2}{132.25} + \frac{(265-264.5)^2}{264.5} + \frac{(126-132.25)^2}{132.25} = 0.546.$$

This statistic follows the chi-square distribution if the null hypothesis is true. The number of degrees of freedom is $k - 1 - m$. In this example, the value of k is 3 for the three types of possible offspring. Since we did not estimate any parameters, m is 0. Therefore, there are 2 degrees of freedom. The critical value, $\chi^2_{2,0.90}$, is 4.61. Since 0.546 is less than 4.61, we fail to reject the null hypothesis. It appears that these data support Mendel's theoretical results.

The goodness-of-fit chi-square statistic can also be used to test the hypothesis that the data follow a particular probability distribution. Thus, it can be used to complement the graphical approaches — for example, the Poissonness and normal probability plots presented in Chapter 5. The test of hypothesis provides a number, the *p*-value, that can be used alone or together with the graphical approach, to help us decide whether or not we will reject or fail to reject the null hypothesis.

Example 10.2

Two Parameters Estimated: Let us test the hypothesis, at the 0.01 significance level, that the systolic blood pressure values for 150 typical 12-year-old boys in the United States, shown in Table 10.2, come from a normally distributed population. In testing the hypothesis that data are from a normally distributed population, we must specify the particular normal distribution. This specification means that the values of the population mean and standard deviation are required. However, since we do not know these values for the systolic blood pressure variable for U.S. 12-year-old boys, we will estimate their values. Table 10.2 shows the sample estimates of the mean and standard deviation.

The goodness-of-fit test is based on the variable of interest being discrete or being grouped into k categories. Therefore, we must group the systolic blood pressures into categories. We use ten categories, shown in Table 10.3.

Table 10.2 Systolic blood pressure values (mmHg) and their sample mean and standard deviation for 150 12-year-old boys.

Value	Freq.	Value	Freq.	Value	Freq.
80	3	100	19	118	2
82	1	102	7	120	7
84	1	104	9	122	2
86	1	105	6	124	2
88	2	106	4	125	3
90	7	108	10	126	2
92	2	110	17	128	2
94	2	112	7	130	5
95	6	114	3	134	2
96	2	115	2	136	1
98	3	116	6	140	2

Sample Mean = 107.45
Sample Standard Deviation = 12.45

Table 10.3 Number of boys observed and expected[a] in the systolic blood pressure categories.

Systolic Blood Pressure (mmHg)	Number	
	Observed	Expected
≤80.5	3	2.28
80.51–87.5	3	5.90
87.51–94.5	13	14.18
94.51–101.5	30	25.10
101.51–108.5	36	32.57
108.51–115.5	29	31.13
115.51–122.5	17	21.83
122.51–129.5	9	11.26
129.51–136.5	7	4.28
≥136.5	3	1.47
Total	150	150.00

[a]Expected calculated assuming the data follow the $N(107.45,12.45)$ distribution

The expected values are found by converting the category boundaries to standard normal values and then finding the probability associated with each category. For example, the probability associated with the first category, a systolic blood pressure less than 80.5 mmHg, is found in the following manner. First, 80.5 is converted to a standard normal value by subtracting the mean and dividing by the standard deviation. Thus, 80.5 is converted to −2.165(= [80.5 − 107.45]/12.45). The probability that a standard normal variable is less than −2.165 is 0.0152. The expected number of observations is found by taking the product of n, 150, and the probability of being in the category. Thus, the expected number of observations in the first category is 2.28 (= 150[0.0152]).

The expected number of observations in the second category is found in the following manner. The upper boundary of the second category, 87.5, is converted to the standard normal value of −1.602 (= [87.5 − 107.45]/12.45). The probability that a standard normal variable is less than −1.602 is 0.0545. The probability of being in the second category is then 0.0393 (= 0.0545 − 0.0152). This probability is multiplied by 150 to get the expected count of 5.90 for the second category. The other expected

cell counts are calculated in this same way. If the sum of the expected counts does not equal the number of observations, with allowance for rounding, an error has been made. Note that three cells have expected counts less than 5. For the chi-square distribution to apply, no cell should have an expected count less than 1.5 (= 5[3/10]). The expected count in the last cell is 1.47, a value that is very close to 1.5. The difference between 1.47 and 1.50 is so slight that we may choose to apply the chi-square distribution, or we could combine the last two categories and use only nine categories in the calculation of the test statistic. Whichever choice we choose should not have much impact on the value of the test statistic.

The calculation of the chi-square goodness-of-fit statistic is

$$X^2 = \frac{(3-2.28)^2}{2.28} + \frac{(3-5.90)^2}{5.90} + \cdots + \frac{(3-1.47)^2}{1.47} = 8.058.$$

The value of k, the number of categories, is 10 and m, the number of parameters estimated, is 2. Therefore, there are 7 degrees of freedom (= 10 − 1 − 2). The value of this test statistic is compared to 18.48 (= $\chi^2_{7,0.99}$). Since 8.058 is less than 18.48, we fail to reject the null hypothesis. Based on this sample, there is no evidence to suggest that the systolic blood pressures of 12-year-old boys are not normally distributed.

In dealing with continuous variables — for example, the blood pressure variable — we have to decide how many intervals and what interval boundaries should be used. In the preceding example, we used 10 intervals. Some research has been conducted on the relation between power considerations and the number and size of intervals, and, as we might expect, the number of intervals depends on the sample size. Table 10.4, based on a review by Cochran (1952), shows the suggested number of intervals to be used with a continuous variable. The size of the intervals may also vary. The intervals can be chosen so that the expected number of observations in each interval is approximately equal. Thus, some intervals may be much narrower than other intervals. For ease of computation, it is reasonable to choose the intervals so that the observed number of observations in each interval is approximately equal. These suggestions for the choice of the number and size of intervals differ from those used in Example 10.2, but the goals of the analyses are also different. In Example 10.2, the equal size intervals were used regardless the number of observations in each interval to determine whether or not it appears that the data follow a particular distribution.

Table 10.4 Guideline for the number of intervals to be used with a continuous variable.

Sample size	200	400	600	800	1000	1500	2000
Number of intervals	15	20	24	27	30	35	39

Source: Cochran, 1952

10.2 The 2 by 2 Contingency Table

In this section, we extend the use of the chi-square goodness-of-fit test statistic to two-way contingency tables. This extension allows a determination of whether or not there is an association between two variables. We begin the study of the association of two

discrete random variables with the simplest two-way table, the 2 by 2 contingency table. This statement by M. H. Doolittle in 1888 expresses the purpose of our analysis:

> The general problem may be stated as follows: Having given the number of instances respectively *in which things are both thus and so*, in which they are *thus but not so*, in which they are *so but not thus*, and in which they are *neither thus nor so*, it is required to eliminate the general quantitative relativity inhering in the mere thingness of the things, and to determine the special quantitative relativity subsisting between the *thusness and the soness* of the things (emphasis added; Goodman and Kruskal 1959).

A restatement of the purpose is that we wish to determine, at some significance level, whether or not there is an association between the variables.

For example, is there is an association between the occurrence of iron deficiency in women and their level of education? If we use two levels of education — for example, less than 12 years and greater than or equal to 12 years — the 2 by 2 table to use in this investigation would look like Table 10.5. The entries in the table, the n_{ij}, are the observed number of women in the ith row (level of education) and jth column (iron status) in the sample. The symbol $n_{i\cdot}$ represents the sum of the frequencies in the ith row, $n_{\cdot j}$ is the sum of the frequencies in the jth column and n, the sample size, is the sum of the frequencies in the entire table.

Table 10.5 Iron status by level of education.

	Iron Status		
Education	**Deficient**	**Acceptable**	**Total**
<12 Years	n_{11}	n_{12}	$n_{1\cdot}$
≥12 Years	n_{21}	n_{22}	$n_{2\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	n

There are several ways of answering the question about whether or not there is an association between these two variables. We begin with the approach from Chapter 7.

10.2.1 Comparing Two Independent Binomial Proportions

The 2 by 2 table is one way of presenting the data used in the calculation of two independent binomial proportions. If there is no association between iron status and education, then the probability of iron deficiency for women with less than 12 years of education, π_1, should equal the corresponding probability, π_2, for women with 12 or more years of education. We can construct a confidence interval for the difference of π_1 and π_2 using the method presented in Chapter 7. If the interval contains zero, there is no evidence of an association between iron status and education. The confidence interval is based on the sample estimates of π_1 and π_2 and these are $n_{11}/n_{1\cdot}$ and $n_{21}/n_{2\cdot}$, respectively.

10.2.2 Expected Cell Counts Assuming No Association: Chi-Square Test

Let us first consider first two common situations when a 2 by 2 table can be formed. The two choices for data collection that are used most often in practice are (1) an SRS

of n observations and (2) stratified samples of $n_{1\cdot}$ and $n_{2\cdot}$ observations. In the SRS case, the test for no association is a test of the independence of the row and column variables. In the stratified sampling case, the test for no association is a test of the homogeneity of the proportions in the ith row with those in the jth row. Regardless of which of these two sample selection processes is used, the expected cell counts for the hypothesis of no association are calculated as shown below.

We use the symbol m_{ij} to represent the expected number of women in the ith row and jth column assuming that the null hypothesis is true. In the material on two-way tables, we are using n and m to represent the observed and expected cell counts instead of the O and E used in the previous section. For the null hypothesis of no association between iron status and education, the expected proportion of women with low iron status at each level of education, $m_{i1}/n_{i\cdot}$, equals the overall proportion of iron deficient women, $n_{\cdot 1}/n$. This is equivalent to saying that the proportion of women with low iron status is the same for those with less than 12 years of education as for those who have at least 12 years of education. Thus, when there is no association, the expected number of iron deficient women at the ith level of education can be found from the following relationship:

$$\frac{m_{i1}}{n_{i\cdot}} = \frac{n_{\cdot 1}}{n}$$

which yields

$$m_{i1} = \frac{n_{i\cdot} n_{\cdot 1}}{n}.$$

The same type of relation holds true for women with acceptable levels of iron. Therefore the general formula for the expected cell count, assuming no association, is

$$m_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n}.$$

We can use these observed and expected values to calculate the chi-square goodness-of-fit statistic to test the hypothesis of no association between the two variables. We often use a modified version of the chi-square goodness-of-fit statistic. The modified form, called the Yates' corrected chi-square after the British statistician, Frank Yates (1984), who suggested it, is

$$X^2_{YC} = \sum_i \sum_j \frac{(|n_{ij} - m_{ij}| - 0.5)^2}{m_{ij}}.$$

The modification consists of subtracting 0.5 from the absolute value of the difference of the observed and expected cell counts. The Yates' corrected chi-square statistic can be calculated directly from the frequencies without calculating the expected counts. The easier-to-use formula is

$$X^2_{YC} = \frac{n(|n_{11}n_{22} - n_{12}n_{21}| - n/2)^2}{n_{1\cdot} n_{2\cdot} n_{\cdot 1} n_{\cdot 2}}.$$

The p-value associated with the Yates' corrected chi-square statistic agrees more closely with the p-value of the exact test statistic developed by Ronald Fisher (1935).

Some statisticians question the use of Fisher's exact test in 2 by 2 tables when the data arise from either of the two sampling methods just discussed. They question the application because Fisher's test was developed based on both the row and column margins being fixed in advance, a different sampling scheme than used in the two methods. Hence, they do not recommend the use of Yates' (1984) correction, but we believe that Yates' correction is appropriate.

Example 10.3

Suppose that we select an SRS of 100 women 20 to 44 years old and we obtain information on their educational level and iron status. The hypothetical data, based on a report (Life Sciences Research Office 1989), are shown in Table 10.6.

The estimated conditional probability of a woman being iron deficient given that she has less than 12 years of education is 0.133 (= 4/30). This is contrasted to the estimated probability of 0.057 (= 4/70) for a woman with 12 or more years of education. Using the procedure in Chapter 7, the 95 percent confidence interval for the difference of π_1 and π_2 is found by

$$(0.133-0.057)\pm z_{0.975}\sqrt{\frac{0.133(1-0.133)}{30}+\frac{0.057(1-0.057)}{70}}$$

which yields an interval from −0.057 to 0.209. Since zero is contained in the interval for the difference, there is no evidence of an association between iron status and education based on this sample.

Based on these data, the expected values, assuming the independence of the row and column variables, are

$$\begin{aligned} m_{11} &= 30\ (8)\ /\ 100 = 2.4,\\ m_{12} &= 30\ (92)\ /\ 100 = 27.6,\\ m_{21} &= 70\ (8)\ /\ 100 = 5.6,\\ m_{22} &= 70\ (92)\ /\ 100 = 64.4,\\ \text{Total} &= 100.0. \end{aligned}$$

The sum of the expected values in the first row is 30, the first row total. The sum of the expected values in the first column is 8, the first column total. Hence, once we calculate m_{11}, we know m_{12}'s value by subtracting m_{11} from 30. In the same way, we know m_{21}'s value by subtracting m_{11} from 8. Since we now know m_{12}'s value, we can also find m_{22}'s value by subtracting m_{12} from 92. Hence, once we calculate any cell's expected value, the expected values of the other three cells are determined.

Table 10.6 Hypothetical frequency data for iron status by education.

	Iron Status		
Education	**Deficient**	**Acceptable**	**Total**
<12 Years	4	26	30
≥12 Years	4	66	70
Total	8	92	100

This means that there is only one degree of freedom associated with the test of no association for a 2 by 2 contingency table.

The expected cell frequency for the cell in the intersection of the first row and first column is 2.4. This is the only expected frequency less than 5, and, according to the guideline just given, the minimum acceptable value for an expected cell frequency is 1.25 (= 5[1/4]). Since none of the expected frequencies are less than 1.25, we can use the chi-square test statistic.

Now that we have both the observed and expected cell counts, we can test the hypothesis of no association (independence) of iron status and education. We shall perform the test at the 0.05 significance level using the Yates' modified chi-square procedure.

The calculated X^2_{YC} is compared to 3.84 (= $\chi^2_{1,0.95}$). If X^2_{YC} is greater than 3.84, we reject the hypothesis of independence in favor of the alternative that there is some association between iron status and education. If X^2_{YC} is less than 3.84, we fail to reject the null hypothesis. The test statistic is

$$X^2_{YC} = \frac{(|4-2.4|-0.5)^2}{2.4} + \frac{(|26-27.6|-0.5)^2}{27.6} + \frac{(|4-5.6|-0.5)^2}{5.6} + \frac{(|66-64.4|-0.5)^2}{64.4} = 0.783.$$

Since X^2_{YC} is less than 3.84, we fail to reject the null hypothesis. Based on this sample, it does not appear that there is any association between iron status and education. Note that the uncorrected X^2 value is 1.656 and we would therefore draw the same conclusion.

The chi-square test can easily be performed using computer packages (see **Program Note 10.1** on the website).

10.2.3 The Odds Ratio — a Measure of Association

A useful statistic for measuring the level of association in contingency tables is the *odds ratio*, θ. For example, in Table 10.5, an estimator of the odds that a woman with less than a high school education is iron deficient is n_{11}/n_{12}. The corresponding estimator of the odds that a woman with at least a high school education is iron deficient is n_{21}/n_{22}. If there is no association between education and iron status, these two odds should be equal. If the odds are equal, their ratio equals one. A sample estimator of the odds ratio *OR* is

$$\hat{\theta} = OR = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{21}n_{12}}.$$

Thus, if *OR* is far from one, it calls into question the assumption (hypothesis) of no association. If the estimated odds ratio is much less than one, this means that the denominator is much larger than the numerator — that is, the product of the off-diagonal cells in the 2 by 2 table is larger than the product of the diagonal cells. For Table 10.5, an odds ratio of less than one indicates that the proportion of women with 12 or more

years of education who are iron deficient is greater than the corresponding proportion for women with fewer than 12 years of education. An odds ratio of greater than one indicates that women with fewer than 12 years of education have the greater proportion of iron deficiency.

A problem with the estimated odds ratio occurs if any of the cell frequencies are zero. The estimated odds ratio is zero if n_{11} or n_{22} are zero, and it is undefined if n_{12} or n_{21} are zero. To avoid this problem, some statisticians base the calculation of the estimated odds ratio on $n_{ij} + 0.5$ instead of the n_{ij}.

We have to realize that there is sampling variation associated with the sample estimate of the odds ratio and this variation must be taken into account in interpreting the estimated odds ratio. Since the distribution of the natural logarithm of θ, $\ln(\theta)$, converges to the normal distribution for smaller sample sizes than the distribution of θ itself, we shall focus on the confidence interval for $\ln(\theta)$. After finding the confidence interval for $\ln(\theta)$, we can transform it to a confidence interval for θ. The estimated standard error for the sample estimate of $\ln(\theta)$ (Agresti 1990) is

$$\hat{\sigma}_{ln(OR)} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}.$$

The $(1 - \alpha) * 100$ percent confidence interval for the $\ln(\theta)$ is

$$\ln(OR) \pm z_{1-\alpha/2}\hat{\sigma}_{ln(OR)}.$$

Example 10.4

The sample odds ratio for the data in Example 10.3 is 2.538 (= 4[66]/4[26]). This value seems to be different from one, and therefore it suggests that there is an association. However, we need to consider its confidence interval.

The estimated standard error for the sample estimate of $\ln(\theta)$ is 0.7441, which is obtained from $\sqrt{1/4 + 1/26 + 1/4 + 1/66}$ The value of the natural logarithm of the sample odds ratio, ln(2.538), is 0.9314. Therefore, the 95 percent confidence interval for $\ln(\theta)$ is $0.9314 \pm 1.96\ (0.7441)$, which ranges from −0.5270 to 2.3897. Taking the exponential of these limits provides the 95 percent confidence interval for θ and its limits are 0.5904 and 10.9104. The confidence interval for the odds ratio is quite large and does include the value of one. Hence, there is no evidence that the null hypothesis should be rejected.

Program Note 10.1 on the website provides a demonstration of these calculations using computer programs.

All three approaches agree that there is no evidence of an association between iron status and education based on this hypothetical sample. These approaches will almost always agree about whether or not an association exists between two variables. The confidence interval for the difference of the probabilities and the uncorrected chi-square statistic will always agree in their conclusions.

10.2.4 Fisher's Exact Test

The chi-square test of association just described relies on the test statistic having an approximate chi-square distribution. This approach is warranted when the expected cell counts are large. However, when very small cell counts (less than 5 times the proportion of cells with expected values less than 5) are involved, the use of chi-square distribution is no longer warranted. Fortunately, an alternative procedure suggested by Fisher is appropriate for such cases.

The basis of Fisher's exact test is to consider all configurations of cell counts given both row and column totals are fixed and to compute the probabilities of the observed configuration and more extreme configurations occurring by chance. If the sum of these probabilities turns out to be very small, we conclude that it is unlikely that such a small value could have occurred by chance, and we then reject the hypothesis of association between the row and column variables. The probability of each configuration is found from the hypergeometric distribution. This probability is based on the number of ways of observing an outcome conditional on fixed margins. To calculate this probability, we must find, using the notation in Table 10.5, the probability of selecting n_{11} elements from $n_{1\cdot}$ and n_{21} elements from $n_{2\cdot}$ given that the row margins are fixed. This probability is found by determining the number of ways selecting n_{11} elements from $n_{1\cdot}$ and n_{21} from $n_{2\cdot}$ and dividing that number by the number of ways selecting n_1 elements from n elements. In symbols, that is

$$\Pr(\textit{configuration}) = \frac{\binom{n_{1\cdot}}{n_{11}}\binom{n_{2\cdot}}{n_{21}}}{\binom{n}{n_{1\cdot}}}$$

which simplifies to

$$\Pr(\textit{configuration}) = \frac{n_{1\cdot}!\,n_{2\cdot}!\,n_{\cdot 1}!\,n_{\cdot 2}!}{n!\,n_{11}!\,n_{12}!\,n_{21}!\,n_{22}!}.$$

Example 10.5

In a small company, the records of promotion in the past year are to be examined for a possible association of gender and promotion. The records show the following:

	Promotion			Percent
Gender	Yes	No	Total	Promoted
Male	5	1	6	83.3
Female	1	4	5	20.0
Total	6	5	11	

We want to test the hypothesis of no association between gender and promotion. We first calculate the expected cell counts under the hypothesis of no association, and they are 3.27 for the (1,1) cell, 2.73 for the (1,2) cell, 2.73 for the (2,1) cell, and 2.27 for the (2,2) cell. To use the chi-square test statistic, we require that none of the

expected values is less than five times the proportion of cells with expectations less than five. In this example, all the expected cell counts are less than five. Therefore, the criterion is also five (= 5[4/4]), and since all the expectations are less than this criterion, we cannot use the chi-square test here. However, we can use Fisher's exact test here. Concentrating on (1,1) cell, we can calculate the probability of observed configuration and a more extreme configuration. The two configurations to be considered are

$$\begin{array}{cc|c} 5 & 1 & 6 \\ 1 & 4 & 5 \\ \hline 6 & 5 & 11 \end{array} \quad \text{and} \quad \begin{array}{cc|c} 6 & 0 & 6 \\ 0 & 5 & 5 \\ \hline 6 & 5 & 11 \end{array}$$

The combined p-value for these two configurations is

$$p-value = \frac{6!5!6!5!}{11!5!1!1!4!} + \frac{6!5!6!5!}{11!5!0!0!5!} = 0.0649 + 0.0022 = 0.0671.$$

The calculated p-value suggests the observed frequencies are somewhat unexpected. However, if we are using the 0.05 level for the test of hypothesis, there is not sufficient evidence to suggest an association between gender and promotion. Even if there were strong evidence of an association, it would not necessarily imply discrimination. There are many other variables that would need to be considered. For example, one important variable would be the date of last promotion. If all the women had been promoted the year before but none of the men, then our interpretation might change.

The calculation of Fisher's exact test statistic for 2 by 2 tables, or for its extension to r by c tables (Mehta and Patel 1983), is quite involved, and the use of computer is recommended. **Program Note 10.1** on the website also includes comments for the Fisher's exact test.

10.2.5 The Analysis of Matched-Pairs Studies

Surprisingly, we can also use the 2 by 2 table in situations with more than two variables. For example, the 2 × 2 table can be used when we wish to determine whether or not there is a relationship between two variables while controlling for other variables. The matched-pair study is one example of this situation.

In the health field we often wish to determine whether or not there is a relationship between disease status and a risk factor while controlling for variables that may affect the relationship. We may have some number of people with some disease of interest (the cases), and we select an equal number of people without the disease (the controls). In an effort to remove the effect of the extraneous variable(s), for each person in the disease group, we pair them with a person from the control group who is the best match on the extraneous variable(s). We present the paired data in a 2 by 2 table as follows where the entries in the table are the observed cell frequencies:

	Control Exposed to Risk Factor	
Case Exposed to Risk Factor	**Yes**	**No**
Yes	c_1	d_1
No	d_2	c_2

Pairs with the same exposure status for both case and control — the diagonal cells — are called concordant pairs ($c_1$and c_2), and pairs with different exposures — the off-diagonal cells — are called discordant pairs (d_1 and d_2).

Let π be the probability that a discordant pair has an exposed case. Then, from the preceding table, π can be estimated by the following proportion,

$$\hat{\pi} = d_1/(d_1 + d_2).$$

Under the null hypothesis of no association between the risk factor and the disease, each discordant pair is just as likely to have a case exposed as to have a control exposed. Thus, the null hypothesis can be written as

$$H_0 : \pi = 1/2.$$

For large samples, we can use the normal approximation as discussed in Chapter 8. In this case the test statistic is

$$z = \frac{d_1/(d_1 + d_2) - (1/2)}{\sqrt{(1/2)(1-1/2)/(d_1 + d_2)}} = \frac{(d_1 - d_2)}{\sqrt{d_1 + d_2}}.$$

Alternatively, we could use the chi-square test with 1 degree of freedom by squaring the z statistic and incorporating Yates's correction for continuity. This chi-square test is referred as McNemar test (1947) for testing no association in a matched study of proportions — that is

$$X_M^2 = \frac{(|d_1 - d_2| - 1)^2}{d_1 + d_2}.$$

We can use this test when $(d_1 + d_2)$ is large. However, this test is not recommended when $(d_1 + d_2)$ is less than 10, since the normal approximation is invalid as discussed in Chapter 5. In that case a preferable testing procedure is the binomial test illustrated in Chapter 5, based on a binominal distribution with $\pi = 0.5$ and $n = (d_1 + d_2)$. The same procedure was used for the sign test in Chapter 9. We also need to pay attention to the size of $(d_1 + d_2)$ in relation to the size of $(c_1 + c_2)$. When concordant pairs are predominant, there is little reason to analyze discordant pairs alone. The McNemar test is an approximate test and should be used with caution for a small data set with a relatively small number of discordant pairs.

Example 10.6

In Chapter 6 we discussed the case-control study design in which people with the disease under investigation (the cases) are compared with people who are free of the disease (the controls). A case-control study of presenile dementia by Forster et al. (1995) identified 109 clinically diagnosed patients aged below 65 years from hospital records. Each case was individually paired with a community control of the same sex and age. Steps were taken to ascertain that the control did not suffer from dementia. One of the risk factors explored in the study was family history of dementia. We wish to determine whether or not there is an association between the occurrence of

presenile dementia and a family history of dementia. The following table shows a crosstabulation of the 109 pairs by the presence or absence of family history of dementia:

Family History of Dementia in Case	Family History of Dementia in Control	
	Present	Absent
Present	6	25
Absent	12	66

Since the cases were paired with the controls, information on the relationship between family history of dementia and disease comes from the 37 discordant pairs. Of these, 25 pairs had an exposed case, twice more than the pairs with an exposed control. The McNemar test statistic is

$$X_M^2 = \frac{(|25-12|-1)^2}{25+12} = 3.89.$$

Compared with 3.84 (= $\chi^2_{1,0.95}$, using Table B7), this is just significant at 0.05 level. Hence, there is evidence for an association between dementia and family history of the disease.

10.3 The *r* by *c* Contingency Table

We now consider the more general situation where two classification variables have more than two categories. First, we consider the situation where both variables are nominal followed by the situation when one of the variables is ordinal.

10.3.1 Testing Hypothesis of No Association

The same ideas used in the 2 by 2 table still apply to the r by c contingency table. If there is no association between a row variable and a column variable, the ratio of the expected cell frequency in the ith row and jth column, m_{ij}, to the ith row total, $n_{i\cdot}$, should equal the ratio of the jth column total, $n_{\cdot j}$, to the overall total. Thus, m_{ij} is still found from

$$\frac{m_{ij}}{n_{i\cdot}} = \frac{n_{\cdot j}}{n}$$

which yields

$$m_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n}.$$

There are $(r-1)(c-1)$ degrees of freedom for the r by c table because once we know the frequencies of any $(r-1)(c-1)$ cells, we can find the values of the other frequencies by subtraction from the row and column totals. The hypothesis of no association between the row and column variables is tested using the chi-square goodness-

of-fit statistic. Most statisticians perform no adjustment to the test statistic when used with tables other than the 2 by 2 table. If the test statistic is greater than the value of $\chi^2_{(r-1)(c-1),1-\alpha}$, we reject the hypothesis of no association in favor of the alternative that the row and column variables are related. If the test statistic is less than $\chi^2_{(r-1)(c-1),1-\alpha}$, we fail to reject the null hypothesis.

Example 10.7

The data in Table 10.7 are from a study in Los Angeles conducted to determine the knowledge and opinion of women about mammography. The study was a response to concern raised in the media about the potential radiation hazards of the long-term use of mammography (Berkanovic and Reeder 1979). Two issues the study addressed were (1) whether or not these articles had caused women to refuse the use of mammography screening for breast cancer and (2) variables related to women's opinions about mammography. A telephone interview was conducted with a sample of women and approximately 60 percent of the women had heard or read something about mammography. Table 10.7 shows the opinion about mammography for those women who had heard or read about it. This is a 2 by 3 table. The question of interest for this table is whether or not there is an association between the woman's opinion about mammography screening and the variable knowing someone with breast cancer. We test this hypothesis at the 0.01 significance level.

There are two (= [2 − 1][3 − 1]) degrees of freedom for this table. Knowing the frequencies for the (1,1) and (1,2) cells allows us to find the value of the (1,3) cell by subtraction of the sum of the (1,1) and (1,2) frequencies from the total of the first row. Knowledge of the frequencies in the first row then allows us to find the cell frequencies in the second row by subtraction from the column totals. For example, the frequency of the (2,1) cell is found by subtracting the frequency of the (1,1) cell from the total of the first column. Similar logic applies to the calculation of expected counts. If we calculate the expected counts for any two cells, then the expected counts for the rest of the cells can be found by subtraction from the row and column totals. The expected counts are also shown in Table 10.7.

The value of the test statistic is found from

$$X^2 = \frac{(120-129.76)^2}{129.76} + \frac{(45-39.52)^2}{39.52} + \ldots + \frac{(8-12.29)^2}{12.29} = 6.65.$$

Table 10.7 Frequency of women by opinion about mammography and whether or not they know someone with breast cancer (expected counts are in parentheses).

	Opinion			
Know Someone with Breast Cancer	**Positive**	**Neutral**	**Negative**	**Total**
Yes	120 (129.76)	45 (39.52)	28 (23.72)	193
No	77 (67.24)	15 (20.48)	8 (12.28)	100
Total	197	60	36	293

Since 6.65 is less than 9.21 ($= \chi^2_{2,0.99}$), we fail to reject the null hypothesis. There does not appear to be a statistically significant association, at the 0.01 level, between opinion about mammography and whether or not someone with breast cancer was known.

We can use the computer to perform the test as shown in **Program Note 10.2** on the website. The *p*-value is (1 − 0.9640) or 0.036. Although a *p*-value of 0.036 is significant at the 0.05 level, the association is not statistically significant at the 0.01 level.

10.3.2 Testing Hypothesis of No Trend

The hypothesis of no association is very general, and it is a reasonable hypothesis to test with nominal variables. However, when a variable conveys more information than the category name, it is possible to test a more specific hypothesis that uses more of the information contained in the variable. For example, in Example 10.5, opinion is an ordinal variable that ranges from positive to neutral to negative, and the test for no association ignores this ordering. In 2 by *c* contingency tables, there is a test, a test for trends that takes the ordering of the column variable into account. There is also a method that can be used for *r* by *c* contingency tables (Semenya et al. 1983).

In the test for no association in Example 10.5, we examined the unconditional cell probabilities. We also could have focused on the conditional probabilities — for example, the probability of women who knew someone with breast cancer conditional on their opinion of mammography. In calculating the conditional probabilities in this fashion, we are not implying that the probability of women who knew someone with breast cancer depends on their opinion of mammography. We are calculating the conditional probabilities in this fashion simply to see if there is a trend in the probabilities of women who knew someone with breast cancer by opinion category. The sample estimates of these conditional probabilities are easily found. For the women who are positive about mammography, the estimated probability of a woman knowing someone with breast cancer is 0.609 (= 120/197). The corresponding values for the women with neutral and negative opinions are 0.750 and 0.778, respectively. If the estimates of these probabilities are related to the opinion category, this suggests that an association exists between the row and column variables.

We now consider the *hypothesis of no linear trend*. By no linear trend, we mean that the proportion of women who knew someone with breast cancer does not increase (decrease) consistently with the changes in opinion from positive to neutral to negative. To perform a test of this hypothesis, we assign a numerical score to the categories of the opinion variable. For example, it seems reasonable to assign a score of +1 to the positive category, 0 to the neutral level and −1 to the negative category. This assignment of scores assumes that the distance from positive to neutral is the same as the distance from neutral to negative. The assignment of scores is subjective and, in unusual situations, the scoring system used can have an impact on the test of hypothesis. However, in most cases, different reasonable scoring systems will lead to the same conclusion about the test of hypothesis.

The hypothesis of no linear trend is basically a test of no correlation between the assigned scores and the conditional probabilities. Thus, the test statistic should look something like a correlation coefficient. The following notation is used in the representation of the test statistic. Let p_j be the sample estimate of the conditional probabilities of women who knew someone with breast cancer and S_j be the score assigned to the jth opinion category, where j equals 1, 2, and 3 for positive, neutral, and negative. The unconditional sample estimate of the women who knew someone with breast cancer is $\bar{p}$, and $\bar{q}$ is $(1 - \bar{p})$. Let $\bar{S}$ be the sample mean score.

The test statistic is

$$X^2 = \frac{\left(\sum_{j=1}^{c} n_{\cdot j}(p_j - \bar{p})(S_j - \bar{S})\right)}{\bar{p}\bar{q}\sum_{j=1}^{c} n_{\cdot j}(S_j - \bar{S})^2}.$$

The numerator of this statistic is the square of the numerator of the correlation coefficient between the conditional proportion and the assigned score. Hence, we can see that this statistic is a measure of the linear trend between these two variables. For sufficiently large sample sizes, this statistic can be shown to follow the chi-square distribution with one degree of freedom if there is no linear trend. The sample size is sufficiently large if, given the value of $\bar{p}$, it is larger than that shown in Table 5.7. Large values of X^2 cause us to reject the null hypothesis of no linear trend in favor of the alternative hypothesis of a linear trend.

Example 10.8

Let us test the null hypothesis of no linear trend in the opinion about mammography data in Example 10.5 at the 0.01 significance level. The overall proportion, $\bar{p}$, of women who knew someone with breast cancer is 0.659 (= 193/293). Hence, $\bar{q}$ is 0.341. Since n is 293, much larger than the values in Table 6.7 for a proportion of 0.30 and 0.35, we can use the test statistic just shown. The p_j are 0.609, 0.750, and 0.778 for j values of 1, 2, and 3. S_1 is +1, S_2 is 0, and S_3 is −1, and the values of the column totals, n_j, are 197, 60, and 36, respectively. The mean of the scores, $\bar{S}$, is found by

$$\frac{197(1) + 60(0) + 36(-1)}{293} = 0.5495.$$

The test statistic is

$$X^2 = \frac{[197(-0.050)(0.4505) + 60(0.091)(-0.5495) + 36(0.119)(1.5495)]^2}{0.659(0.341)\left[197(0.4505)^2 + 60(-0.5495)^2 + 36(-1.5495)^2\right]}$$

$$= \frac{(-14.076)^2}{32.479} = 6.100.$$

This statistic is compared to 6.63 (= $\chi^2_{1,0.99}$). Since 6.100 is less than 6.63, we fail to reject the null hypothesis of no linear trend. The p-value of this test statistic is found to be 0.0135. Although there is not a statistically significant linear trend in these data at the 0.01 significance level, there is a strong inverse relationship between

the conditional proportion of women who knew someone with breast cancer and their opinion about mammography. We know the relationship is inverse because the sign of the numerator, before squaring, is negative. The opinion about mammography is more likely to be neutral or negative as the proportion of women who knew someone with breast cancer increases.

For the use of computer for the preceding analysis, see **Program Note 10.3** on the website.

This test for trends is equivalent to creating a confidence interval for the difference in means from two independent populations. In this example, the two independent populations are the women who did not know someone with breast cancer and those who did know someone with breast cancer.

The test for trends is particularly appropriate for 2 by c contingency tables when there is an ordering among the column categories. If a linear trend exists, it may be missed by the general test for association, whereas the trend test has a greater chance of detecting it. The general test for association could cause us to say that there is no relationship between the rows and columns when there actually was a linear trend.

10.4 Multiple 2 by 2 Contingency Tables

Most studies involve the analysis of more than two variables at one time. Often we are interested in the relation between an independent variable and the dependent variable, but there is an extraneous variable that must also be considered. For example, consider a study to determine if there is any association between the occurrence of upper respiratory infections (URI) of young children and outdoor air pollution. There are several variables that could affect the relationship between the occurrence of infections and outdoor air pollution. One variable is the quality of the indoor air. One easily obtained variable that partially addresses the indoor air quality is whether or not someone smokes in the home. This variable is likely to be related to the dependent variable, the occurrence of URI, and it may also be related to the independent variable. Hypothetical data for this situation are based on an article by Jaakkola et al. (1991) and are shown in Table 10.8.

Table 10.8 Number of 6-year-old Finnish children by respiratory status and pollution level with a control for passive smoke in the home.[a]

		Upper Respiratory Infection during Previous 12 Months		
Passive Smoke in the Home	**City Polluted**	**Some**	**None**	**Total**
Yes	High	100	20	120
	Low	124	40	164
	Total	224	60	284
No	High	128	62	190
	Low	166	119	285
	Total	294	181	475

[a]The entries in the table are based on an article by Jaakkola, but the data are hypothetical.

10.4.1 Analyzing the Tables Separately

If we ignore the passive smoke variable, the X^2_{YC} for the combined table is 6.387, its p-value is 0.0115, and the estimate of the odds ratio is 1.524. There is a statistically significant relationship at the 0.05 level between the outdoor pollution variable and the occurrence of URI. The estimated odds ratio of 1.524 means that the odds of URI during the previous 12 months is about 1½ times greater in the city with high pollution than in the city with low pollution. However, this analysis has excluded the passive smoke variable, a variable that we wish to take into account.

One way of taking the passive smoke variable into account is to analyze each 2 by 2 table separately. In this example, the X^2_{YC} statistic is 2.039 and its p-value is 0.1533 for homes in which someone smoked. The X^2_{YC} value is 3.645, and its p-value is 0.0562 for those without passive smoke in the home.

The corresponding estimates of the odds ratios for these two tables are 1.613 and 1.480. The 95 percent confidence intervals for the two odds ratios are from 0.887 to 2.933 and from 1.007 to 2.171, respectively. The first confidence interval, a much wider interval than the second interval, includes the value of one that suggests that there is no relation between the two variables. The second interval barely misses including one. The second interval's smaller size reflects the larger sample size associated with the home in which there was no passive smoke. Neither of these tables has a statistically significant association between the outdoor air pollution and the occurrence of URI at the 0.05 level based on the test statistics. The conclusion from the analyses of the separate tables is different from that of the combined table.

A problem with the use of the separate tables is that the analyses are based on the smaller sample sizes associated with each subtable, not on the sample size of the combined table. This makes it difficult to find the presence of small but consistent trends across tables. A method for eliminating this problem is discussed in the next section. However, before presenting the method, we should also consider a problem that can occur when subtables are combined.

Besides ignoring the passive smoke variable, a potential problem in using the combined table is that it can be misleading. For example, if the data are selected from a population that does not represent the target population, strange things can occur. Suppose that we want our results to apply to all children in Finland but that the children used in this study were sampled from those who had been hospitalized during the previous 12 months. If this were done, the population used in the study would not match the target population. Is that a problem? As we have said before, the decision on the generalizability of the results to the target population depends on substantive considerations, not on statistical ideas. Let us assume that the sample data are those in Table 10.9.

In both of the subtables, the city with the lesser pollution had the greater proportion of children with no URI during the past 12 months. If we ignore the passive smoke variable, the combined table is Table 10.10.

In the combined table, the city with the greater outdoor pollution now has the greater proportion of children with no URI during the past 12 months — 0.624 compared to 0.595 for the city with lesser pollution. This example points out that care must be exercised in combining tables when the population from which the sample was drawn is not

Table 10.9 Number of 6-year-old Finnish children by respiratory status and pollution level with a control for passive smoke in the home based on taking samples from a list of hospitalized children.[a]

		Upper Respiratory Infection during Previous 12 Months		
Passive Smoke in the Home	**City Polluted**	**Some**	**None**	**Total**
Yes	High	35	40	75
	Low	60	80	140
	Total	95	120	215
No	High	170	300	470
	Low	15	30	45
	Total	185	330	515

[a]Hypothetical data

Table 10.10 Number of children with occurrence of upper respiratory infection by pollution status of city ignoring the passive smoke variable.

	Upper Respiratory Infection during Previous 12 Months		
City Polluted	**Some**	**None**	**Total**
High	205	340	545
Low	75	110	185
Total	280	450	730

representative of the target population. This was clearly pointed out in an article by Berkson in 1946.

10.4.2 The Cochran-Mantel-Haenszel Test

Two biostatisticians, Nathan Mantel and William Haenszel, developed a method in 1959 for examining the relation between two categorical variables while controlling for another categorical variable (Mantel and Haenszel 1959). This method, similar to a method published by William Cochran in 1954, uses all the data in the combined table and produces one overall test statistic. The test is designed to detect the consistent effect of the independent variable on the dependent variable across the levels of the extraneous variable. Thus, this method should only be used when the estimated odds ratios in the subtables are similar to one another. One very attractive feature of this test is that it can be used with extremely small sample sizes. This test has also been generalized for application to three-way tables of size other than 2 by 2 by k (Landis, Heyman, and Koch 1978).

To facilitate the presentation of the test statistic, we shall use the following notation for the ith 2 by 2 contingency table, where i ranges from one to k, the number of levels of the extraneous variable. In our example, k is 2, since there are only two levels, presence or absence, of the passive smoke variable. The ith 2 by 2 table is shown here.

	Upper Respiratory Infection		
Polluted City	**Some**	**None**	**Total**
High	a_i	b_i	$a_i + b_i$
Low	c_i	d_i	$c_i + d_i$
Total	$a_i + c_i$	$b_i + d_i$	n_i

The test statistic is based on an overall comparison of the observed and expected in the (1,1) cell in each of the k subtables. As we saw earlier in this chapter, under the hypothesis of no association between the row and column variables, there is only one degree of freedom associated with the table. Hence, we key on only one cell in the table and the choice of which cell is arbitrary. A statistic that could be used to examine whether or not there is an association is

$$Z^* = \frac{\sum_{i=1}^{k}(O_i - E_i)}{s.e.\left[\sum_{i=1}^{k}(O_i - E_i)\right]}$$

where O_i and E_i are the observed and expected values in the (1,1) cell in the ith subtable. This statistic is very similar to a standard normal variable where E_i is analogous to the hypothesized mean in the standard normal variable.

In terms of the entries in the ith table, E_i is defined as

$$E_i = \frac{(a_i + b_i)(a_i + c_i)}{n_i}$$

the product of the row total and the column total divided by the table's sample size. The observed (1,1) cell frequency, O_i, is a_i. V_i, with a variance of O_i minus E_i, can be shown to be

$$V_i = \frac{(a_i + b_i)(c_i + d_i)(a_i + c_i)(b_i + d_i)}{n_i^2\,(n_i - 1)}.$$

Because we are dealing with discrete variables, we should use the continuity correction with Z*. However, instead of using the continuity-corrected Z* statistic, we would prefer to use a chi-square statistic, since all the other tests associated with contingency tables use a chi-square statistic. This poses no problem, since the square of a standard normal variable follows a chi-square distribution with one degree of freedom. Thus, the statistic to be used to test the hypothesis of no association between air pollution and the occurrence of upper respiratory problems is the Cochran-Mantel-Haenszel chi-square statistic. Also called the Mantel-Haenszel statistic, it is defined by

$$X^2_{CMH} = \frac{(|O - E| - 0.5)^2}{V}$$

where O, E, and V are defined as the sums of the O_i, the E_i and the V_i over the k subtables. If X^2_{CMH} is greater than $\chi^2_{1,1-\alpha}$, we reject the hypothesis of no association between air pollution and the occurrence of upper respiratory infections. Otherwise we fail to reject the null hypothesis.

Example 10.9

Let us apply this method to the data in Table 10.8. Since the odds ratios in the two separate subtables were similar — 1.613 in homes with passive smoke and 1.480 in the other homes — we can use the X^2_{CMH} statistic. If the odds ratios had not been similar, the effect of the independent variable on the dependent variable is not consistent across the levels of the extraneous variable. Hence, it would not make sense to use the CMH statistic to test for a consistent effect of the independent variable when we already know that such an effect does not exist. Since the values of our odds ratios are similar, we can test the hypothesis of no association (no consistent effect) of air pollution with the occurrence of URI while controlling for passive smoke status, and we shall perform the test at the 0.05 significance level. From Table 10.8 we see that O_1 is 100 and O_2 is 128 and their sum is 228. The expected values are calculated to be

$$E_1 = \frac{120(224)}{284} = 94.65 \quad \text{and} \quad E_2 = \frac{190(294)}{475} = 117.60$$

and their sum is 212.25. The variances are calculated to be

$$V_1 = \frac{120(164)(60)(224)}{284^2\,(283)} = 11.59 \quad \text{and} \quad V_2 = \frac{190(285)(181)(294)}{475^2\,(474)} = 26.94$$

and their sum is 38.53. Thus we have the pieces needed to calculate X^2_{CMH} and its value is

$$X^2_{CMH} = \frac{(|228 - 212.26| - 0.5)^2}{38.53} = 6.036.$$

Since 6.036 is greater than 3.84 (= $\chi^2_{1,0.95}$), we reject the hypothesis of no association. At the 0.05 level, we conclude that there is an association between air pollution and URI even after controlling for passive smoke in the home.

10.4.3 The Mantel-Haenszel Common Odds Ratio

Mantel and Haenszel also showed how to combine the data from the separate subtables to form a common odds ratio for the data. Again this should only be done when the estimated odds ratios in the subtables are similar. If the estimated odds ratios for the subtables are not similar — for example, some are less than one and some are greater than one — the common odds ratio would not be very useful. The relation between the independent and dependent variable would depend on the level of the extraneous variable, and the use of a common odds ratio would mask this. The Mantel-Haenszel estimator of the common odds ratio, θ, is

$$OR_{MH} = \frac{\sum_{i=1}^{k} [a_i\,(d_i / n_i)]}{\sum_{i=1}^{k} [b_i\,(c_i / n_i)]}.$$

There are several approaches to finding an estimate of the variance of the Mantel-Haenszel estimator of the common odds ratio (Letters 1993; Mehta and Walsh 1992), but they are quite involved and will not be presented here.

Example 10.10

For the air pollution data in Table 10.8, the Mantel-Haenszel estimate of common odds ratio is found from

$$OR_{MH} = \frac{[100(40/284)] + [128(119/475)]}{[20(124/284)] + [62(166/475)]} = 1.517.$$

This value is similar to the individual odds ratios of 1.613 and 1.480 and also close to the value, 1.524, found from the overall table. The similarity of the values supports the finding that the passive smoke variable had little effect on the relation between air pollution and URI.

Program Note 10.4 on the website shows the commands used to perform the calculation shown in Examples 10.7 and 10.8. The standard error and confidence intervals for the common odds ratio are also provided.

Conclusion

In this chapter, we introduced another nonparametric test — the chi-square goodness-of-fit test — and showed its use with one- and two-way contingency tables. We also showed two related methods — comparison of two binomial proportions and the calculation of the odds ratio — for determining, at some significance level, whether or not there is a relation between two discrete variables with two levels each. The odds ratio is of particular interest as it is used extensively in epidemiologic research. We also presented the extension of the goodness-of-fit test for no interaction to r by c contingency tables. Another test shown was the trend test, and it is of interest because it has a greater chance of detecting a linear relationship between a nominal and an ordinal variable than does the general chi-square test for no interaction. We also showed different ways for testing the hypothesis of no relationship between two discrete variables with two levels each in the matched-pairs situation. The Cochran-Mantel-Haenszel test and estimate of the common odds ratio were introduced for multiple 2 by 2 contingency tables. These procedures are also used extensively by epidemiologists. In the next chapter, we conclude the material on nonparametric procedures with the presentation of several nonparametric methods for the analysis of survival data.

EXERCISES

10.1 The following data are from one of the hospitals that participated in a study performed by the Veterans Administration Cooperative Duodenal Ulcer Study Group (Grizzle, Starmer, and Koch 1969). The data from 148 men show the severity of an undesirable side effect, the dumping syndrome, of surgery for duodenal ulcer for four surgical procedures. The procedures are the following:

A is drainage and vagotomy; B is 25 percent resection (antrectomy) and vagotomy; C is 50 percent resection (hemigastrectomy) and vagotomy; and D is 75 percent resection.

	Severity of Dumping Syndrome			
Surgery	**None**	**Slight**	**Moderate**	**Total**
A	23	7	2	32
B	23	10	5	38
C	20	13	5	38
D	24	10	6	40
Total	90	40	18	148

Was the design used in this hospital a completely randomized design or a randomized block design? Explain your answer. Test the hypothesis of no association between the type of surgery and the severity of the side effect at the 0.05 significance level. Assuming that the procedures are equally effective, would you recommend any of the procedures over the others?

10.2 Test the hypothesis that the data from Gosset, shown in Table 5.4 and repeated here, come from a Poisson distribution at the 0.01 significance level (Poisson probabilities are shown in Table 5.4).

Observed Frequency of Yeast Cells in 400 Squares								
X	0	1	2	3	4	5	6	Total
Frequency	103	143	98	42	8	4	2	400

10.3 The following data, from an article by Cochran (1954), show the clinical change by degree of infiltration — a measure of a type of skin damage — present at the beginning of the study for 196 leprosy patients who received 48 weeks of treatment.

Degree of	**Improvement**					
Infiltration	**Worse**	**Same**	**Slight**	**Moderate**	**Marked**	**Total**
0–7	11	27	42	53	11	144
8–15	7	15	16	13	1	52
Total	18	42	58	66	12	196

Test the hypothesis of no association between the degree of infiltration and the clinical change at the 0.05 significance level. Is this a test of independence or homogeneity? Explain your answer. Now assign scores from –1 to +3 for the clinical change categories worse to marked improvement and test the hypothesis of no linear trend at the 0.05 significance level. Is there any difference in the results of the tests? Select another reasonable set of scores and perform the trend test again using the second set of scores. Is the result consistent with the result from the first set of scores?

10.4 Mantel (1963) provided data from a study to determine whether or not there is any difference in the effectiveness of immediately injecting or waiting 90

minutes before injecting penicillin in rabbits who have been given a lethal injection. An extraneous variable is the level of penicillin. The data are shown in the following table.

		Response		
Penicillin Level	**Delay**	**Cured**	**Died**	**Total**
1/8	None	0	6	6
	90 minutes	0	5	5
1/4	None	3	3	6
	90 minutes	0	6	6
1/2	None	6	0	6
	90 minutes	2	4	6
1	None	5	1	6
	90 minutes	6	0	6
4	None	2	0	2
	90 minutes	5	0	5

Is it appropriate to use the CMH statistic here to test the hypothesis of no association between the delay and response variables while controlling for the penicillin level? Explain your answer. If you feel that it is appropriate to use the CMH statistic here, test, at the 0.01 significance level, the hypothesis of no association between the delay and response variables while controlling for the penicillin level.

10.5 Your local health department conducts a course on food handling. To evaluate this course, you select an SRS of restaurants from the list of licensed restaurants. For these restaurants in your sample, you then search the list of violations found by the health department during the last two years as well as the list of restaurants with employees who have attended the course during the last two years. For the 86 sampled restaurants, the data can be presented as follows:

	Violation		
Attended Course	**Yes**	**No**	**Total**
Yes	9	28	37
No	36	13	49
Total	45	41	86

Use an appropriate procedure to test the hypothesis of no association between course attendance and the occurrence of a violation at the 0.10 significance level.

Based on these data, discuss whether or not course attendance had any effect on the finding of a restaurant's violation of the health code.

10.6 Cochran (1954) presented data on erythroblastosis foetalis, a sometimes fatal disease in newborn infants. The disease is caused by the presence of an anti-*Rh* antibody in the blood of an *Rh*+ baby. One treatment used for this disease is the transfusion of blood that is free of the anti-*Rh* antibody. In 179 cases in which this treatment was used in a Boston hospital, there were no infant deaths out of 42 cases when the donor was female compared to 27 deaths when the donor was male. One possible explanation for this surprising finding was that the male donors were used in the more severe cases. Therefore, the disease

severity was taken into account and the data are shown in the following table:

		Survival Status		
Disease Severity	**Donor's Sex**	**Dead**	**Alive**	**Total**
None	M	2	21	23
	F	0	10	10
Mild	M	2	40	42
	F	0	18	18
Moderate	M	6	33	39
	F	0	10	10
Severe	M	17	16	33
	F	0	4	4
Total		27	152	179

Use the CMH statistic to test the hypothesis of no association between donor's sex and the survival status of the infant at the 0.05 significance level.

10.7 Group the blood pressure values shown in Table 10.2 into categories of <80, 80–89, 90–99, 100–109, 110–119, 120–129, ≥130 mmHg. Based on this grouping, test the hypothesis that the systolic blood pressure of 12-year-old boys follows a normal distribution using the 0.05 significance level. Compare your results to those based on the grouping shown in Table 10.3.

10.8 The following data show the relation between two types of media exposure and a person's knowledge of cancer (Forthofer and Lehnen 1981).

Media Exposure		Knowledge of Cancer	
Newspapers	**Radio**	**Good**	**Poor**
Read	Listen	168	138
	Do not listen	310	357
Do not read	Listen	34	72
	Do not listen	156	494
Total		668	1061

Based on these data, test the hypothesis of no association between newspapers and knowledge of cancer, ignoring the radio variable. Next, test the hypothesis of no association between radio and knowledge of cancer, ignoring the newspaper variable. Which variable has the stronger association with the knowledge of cancer variable? Based on these data, would you feel comfortable recommending one of these media over the other for the purpose of increasing the public's knowledge of cancer? If your answer is yes, what assumptions are you making about the data? If your answer is no, provide your rationale for your answer.

10.9 Two pathologists each examined coded material from the same 100 tumors and classified the material as malignant or benign. Pathologist A found that 18 are malignant, and pathologist B found 10 malignant cases. Both pathologists agreed on 8 cases as malignant. The investigator conducting the study is interested in determining the extent to which the pathologists differ in their assessments of the tumor material. Form an appropriate 2 by 2 table and test the null hypothesis of no difference at the 0.05 significance level.

REFERENCES

Agresti, A. *Categorical Data Analysis*. John Wiley & Sons, Inc., 1990, p. 54–55.

Berkanovic, E., and S. J. Reeder "Awareness, Opinion and Behavioral Intention of Urban Women Regarding Mammography." *American Journal of Public Health* 69:1172–1174, 1979.

Berkson, J. "Limitations of the Application of Fourfold Table Analysis to Hospital Data." *Biometrics Bulletin* (now *Biometrics*) 2:47–53, 1946.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland. *Discrete Multivariate Analysis: Theory and Practice*. Boston, MA: The MIT Press, 1975, p. 328.

Cochran, W. G. "The χ^2 Test of Goodness of Fit." *Annals of Mathematical Statistics* 23:315–345, 1952.

——— "Some Methods for Strengthening the Common χ^2 Tests." *Biometrics* 10:417–451, 1954.

Fisher, R. A. "The Logic of Inductive Inference (with Discussion)." *Journal of the Royal Statistical Society* 98:39–82, 1935.

Forster, D. P., A. J. Newens, D. W. K. Kay, and J. A. Edwardson. "Risk Factors in Clinically Diagnosed Presenile Demential of the Alzheimer Type: A Case-Control Study in Northern England." *Journal of Epidemiology and Community Health* 49: 253–258, 1995.

Forthofer, R. N., and R. G. Lehnen. *Public Program Analysis: A New Categorical Data Approach*. Belmont, CA. Lifetime Learning Publications, 1981, p. 36.

Goodman, L. A., and W. H. Kruskal. "Measures of Association for Cross-Classifications, II: Further Discussion and References." *Journal of the American Statistical Association* 54:123–163, 1959, p. 131.

Grizzle, J. E., C. F. Starmer, and G. G. Koch. "Analysis of Categorical Data for Linear Models." *Biometrics* 25:489–504, 1969.

Jaakkola, J. J. K., M. Paunio, M. Virtanen, and O. P. Heinonen. "Low-Level Air Pollution and Upper Respiratory Infections in Children." *American Journal of Public Health* 81:1060–1063, 1991.

Landis, J. R., E. R. Heyman, and G. G. Koch. "Average Partial Association in Three-Way Contingency Tables: A Review and Discussion of Alternative Tests." *International Statistical Review* 46:237–254, 1978.

Letters to the Editor. *The American Statistician* 47:86–87, 1993.

Life Sciences Research Office, Federation of American Societies for Experimental Biology: *Nutrition Monitoring in the United States — An Update Report on Nutrition Monitoring*. Prepared for the U.S. Department of Agriculture and the U.S. Department of Health and Human Services. DHHS Pub. No. (PHS) 89–1255, 1989, Figure 6–13.

Mantel, N. "Chi-Square Tests with One Degree of Freedom; Extensions of the Mantel-Haenszel Procedure." *Journal of the American Statistical Association* 58:690–700, 1963.

Mantel, N., and W. Haenszel. "Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease." *Journal of the National Cancer Institute* 22:719–748, 1959.

McNemar, Q. "Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages." *Psychometrika* 12:153–157, 1947.

Mehta, C. R., and N. R. Patel. "A Network Algorithm for Performing Fisher's Exact Test in r × c Contingency Tables." *Journal of the American Statistical Association* 78:427–434, 1983.

Mehta, C. R., and S. J. Walsh. "Comparison of Exact, Mid-p, and Mantel-Haenszel Confidence Intervals for the Common Odds Ratio Across Several 2 × 2 Contingency Tables." *The American Statistician* 46:146–150, 1992.

Semenya, K. A., G. G. Koch, M. E. Stokes, and R. N. Forthofer. "Linear Models Methods for Some Rank Function Analyses of Ordinal Categorical Data." *Communications in Statistics* 12:1277–1298, 1983.

Snyder, L. H. "Heredity." In *Collier's Encyclopedia* 12:68–76, 1970.

Yarnold, J. K. "The Minimum Expectation in χ^2 Goodness of Fit Tests and the Accuracy of Approximations for the Null Distribution." *Journal of the American Statistical Association* 65:864–886, 1970.

Yates, F. "Tests of Significance for 2×2 Contingency Tables (with Discussion)." *Journal of the Royal Statistical Society A* 147:426–463, 1984.