CHAPTER 2

Two-Way Contingency Tables

Table 2.1 cross classifies a sample of Americans according to their gender and their opinion about an afterlife. For the females in the sample, for instance, 435 said they believed in an afterlife and 147 said they did not or were undecided. For such data, we might study whether an association exists between gender and belief in an afterlife. Is one sex more likely than the other to believe in an afterlife, or is belief in an afterlife independent of gender?

Analyzing associations is at the heart of most multivariate statistical analyses. This chapter deals with associations between two categorical variables. We introduce parameters that describe the association and present inferential methods for those parameters.

Many applications involve comparing two groups with respect to the relative numbers of observations in two categories. For Table 2.1, one might compare the proportions of males and females who believe in an afterlife. For such data, Section 2.2 presents methods for analyzing differences and ratios of proportions. Section 2.3 presents another measure, the *odds ratio*, that plays a key role for several methods discussed in this text. Sections 2.4 and 2.5 describe large-sample significance tests about whether an association exists between two categorical variables; Section 2.4 presents tests for nominal variables, and Section 2.5 presents an alternative test for ordinal variables. Section 2.6 discusses small-sample analyses. First, Section 2.1 introduces terminology and notation.

2.1 PROBABILITY STRUCTURE FOR CONTINGENCY TABLES

Categorical data consist of frequency counts of observations occurring in the response categories. Let X and Y denote two categorical variables, X having I levels and Y having J levels. We display the IJ possible combinations of outcomes in a rectangular table having I rows for the categories of X and J columns for the categories of Y. The *cells* of the table represent the IJ possible outcomes. A table of this form in

Table 2.1 Cross Classification of Belief in Afterlife by Gender

	Belief in Afterlife	
Gender	Yes	No or Undecided
Females	435	147
Males	375	134

Source: Data from 1991 General Social Survey.

which the cells contain frequency counts of outcomes is called a *contingency table*. A contingency table that cross classifies two variables is called a *two-way table*; one that cross classifies three variables is called a *three-way table*, and so forth. A two-way table having I rows and J columns is called an $I \times J$ (read I-by-J) table. Table 2.1, for instance, is a 2×2 table.

2.1.1 Joint, Marginal, and Conditional Probabilities

Probability distributions for contingency tables relate to the sampling scheme, as we shall discuss in Section 2.1.4. We first present the fundamental types of probabilities for two-way contingency tables. Suppose first that each subject in a sample is randomly chosen from some population of interest, and then classified on two categorical responses, X and Y. Let $\pi_{ij} = P(X = i, Y = j)$ denote the probability that (X, Y) falls in the cell in row i and column j. The probabilities $\{\pi_{ij}\}$ form the *joint distribution* of X and Y. They satisfy $\sum_{i,j} \pi_{ij} = 1$.

The *marginal distributions* are the row and column totals of the joint probabilities. These are denoted by $\{\pi_{i+}\}$ for the row variable and $\{\pi_{+j}\}$ for the column variable, where the subscript "+" denotes the sum over the index it replaces. For instance, for 2×2 tables,

$$\pi_{1+} = \pi_{11} + \pi_{12} \quad \text{and} \quad \pi_{+1} = \pi_{11} + \pi_{21}.$$

We use similar notation for samples, with Roman p in place of Greek π. For instance, $\{p_{ij}\}$ denotes the sample joint distribution. These are the sample cell proportions. The cell counts are denoted by $\{n_{ij}\}$, with $n = \sum_{i,j} n_{ij}$ denoting the total sample size. The cell proportions and cell counts are related by

$$p_{ij} = \frac{n_{ij}}{n}.$$

The marginal frequencies are the row totals $\{n_{i+}\}$ and the column totals $\{n_{+j}\}$.

In many contingency tables, one variable (say, the column variable, Y) is a response variable and the other (the row variable, X) is an explanatory variable. Then it is informative to construct a separate probability distribution for Y at each level of X. Such a distribution consists of *conditional probabilities* for Y, given the level of X, and is called a *conditional distribution*.

Table 2.2 Notation for Table 2.1

	Belief in Afterlife		
Gender	Yes	No or Undecided	Total
Females	$n_{11} = 435$	$n_{12} = 147$	$n_{1+} = 582$
Males	$n_{21} = 375$	$n_{22} = 134$	$n_{2+} = 509$
Total	$n_{+1} = 810$	$n_{+2} = 281$	$n = 1091$

2.1.2 Belief in Afterlife Example

Table 2.1, a 2×2 contingency table, cross classifies $n = 1091$ respondents to the 1991 General Social Survey by their gender and their belief in an afterlife. Table 2.2 illustrates the cell count notation for these data. For instance, $n_{11} = 435$, and the corresponding sample joint proportion is $p_{11} = 435/1091 = .399$.

In Table 2.1, belief in the afterlife is a response variable and gender is an explanatory variable. We therefore study the conditional distributions of belief in the afterlife, given gender. For females, the proportion of "yes" responses was $435/582 = .747$, and the proportion of "no" responses was $147/582 = .253$. The proportions $(.747, .253)$ are females' sample conditional distribution of belief in the afterlife. For males, the sample conditional distribution is $(.737, .263)$. (Problem 2.11 requests further analyses of these data using methods of this chapter.)

2.1.3 Independence

Two variables are said to be *statistically independent* if the conditional distributions of Y are identical at each level of X. When two variables are independent, the probability of any particular column response j is the same in each row. Belief in an afterlife is independent of gender, for instance, if the actual probability of believing in an afterlife equals .740 both for females and for males.

When both variables are response variables, one can describe their relationship using their joint distribution, or the conditional distribution of Y given X, or the conditional distribution of X given Y. Statistical independence is, equivalently, the property that all joint probabilities equal the product of their marginal probabilities,

$$\pi_{ij} = \pi_{i+}\pi_{+j} \quad \text{for } i = 1, \ldots, I \quad \text{and} \quad j = 1, \ldots, J.$$

That is, the probability that X falls in row i and Y falls in column j is the product of the probability that X falls in row i with the probability that Y falls in column j.

2.1.4 Poisson, Binomial, and Multinomial Sampling

The sampling models introduced in Section 1.2 extend to cell counts in contingency tables. For instance, the Poisson sampling model for a 2×2 table treats each of the four cell counts in the table as an independent Poisson variate.

When the rows of a contingency table refer to different groups, the sample sizes for those groups are often fixed by the sampling design. The next section contains a 2×2 table of this type. The first row refers to 11,000 subjects receiving one treatment, the second row refers to 11,000 separate subjects receiving a different treatment, and each subject is measured on a categorical response variable. When the marginal totals for the levels of X are fixed rather than random, a joint distribution for X and Y is no longer meaningful, but conditional distributions for Y at each level of X are. When there are two response categories, we assume a binomial distribution for the sample in each row, with number of trials equal to the fixed row total. When the samples in the rows are independent, such as separate random samples, this sampling scheme is called *independent binomial sampling*.

When the total sample size in the table is fixed but not the row or column totals, a *multinomial* sampling model applies, in which the cells are the possible outcomes. For instance, Table 2.1 cross classifies a random sample of 1091 subjects according to gender and belief in afterlife. The four cell counts are sample values from a multinomial distribution having four categories.

For many multinomial samples over the cells of a contingency table, the columns are a response variable and the rows are an explanatory variable. Then, to describe the data, it is sensible to divide the cell counts by the row totals to form conditional distributions on the response. In doing so, we inherently treat the row totals as fixed and analyze the data the same way as if they formed separate independent samples. In Table 2.1, for instance, we might treat the results for females as a binomial sample with outcome categories "yes" and "no or undecided" for belief in an afterlife, and the results for males as an independent binomial sample on the same response variable. If there were more than two response categories, such as ("yes," "no," "undecided"), we would treat the samples as independent multinomial samples.

For most analyses, one need not worry about which sampling model makes the most sense. For the primary inferential methods in this text, the same results occur for Poisson, multinomial, and independent binomial/multinomial sampling models.

2.2 COMPARING PROPORTIONS IN TWO-BY-TWO TABLES

Response variables having two categories are called *binary variables*. For instance, "belief in afterlife" is binary when measured using the categories (yes, no). Many studies compare two groups on a binary response, Y. The data can be displayed in a 2×2 contingency table, in which the rows are the two groups and the columns are the response levels of Y. This section presents measures for comparing two groups on binary responses.

2.2.1 Difference of Proportions

We use the generic terms *success* and *failure* for the response categories of a binary variable. For subjects in row 1, let π_1 denote the probability of a success, with $1 - \pi_1$ the probability of a failure. The probabilities $(\pi_1, 1 - \pi_1)$ form the conditional

distribution of Y in row 1. For subjects in row 2, let π_2 denote the probability of success.

The *difference of proportions* $\pi_1 - \pi_2$ compares the success probabilities in the two rows. This difference falls between -1 and $+1$. It equals zero when $\pi_1 = \pi_2$; that is, when the response is independent of the group classification.

Let p_1 and p_2 denote *sample* proportions of successes for the two rows. The sample difference $p_1 - p_2$ estimates $\pi_1 - \pi_2$. For instance, in row 1 of Table 2.2, $p_1 = n_{11}/n_{1+} = 435/582 = .747$ is the number of "yes" responses divided by the sample size in that row; $p_2 = 375/509 = .737$ is the corresponding sample proportion in row 2. The sample difference of proportions is $.747 - .737 = .010$.

For simplicity, we denote the sample sizes for the two groups (i.e., the row totals n_{1+} and n_{2+}) by N_1 and N_2. When the counts in the two rows are independent binomial samples, the estimated standard error of $p_1 - p_2$ is

$$\hat{\sigma}(p_1 - p_2) = \sqrt{\frac{p_1(1-p_1)}{N_1} + \frac{p_2(1-p_2)}{N_2}}. \tag{2.2.1}$$

The standard error decreases, and hence the estimate of $\pi_1 - \pi_2$ improves, as the sample sizes increase. A large-sample $100(1-\alpha)\%$ confidence interval for $\pi_1 - \pi_2$ is

$$(p_1 - p_2) \pm z_{\alpha/2}\,\hat{\sigma}(p_1 - p_2), \tag{2.2.2}$$

where $z_{\alpha/2}$ denotes the standard normal percentile having right-tail probability equal to $\alpha/2$ (e.g., for a 95% interval, $\alpha = .05$, $z_{\alpha/2} = z_{.025} = 1.96$).

2.2.2 Aspirin and Heart Attacks Example

Table 2.3 is taken from a report on the relationship between aspirin use and myocardial infarction (heart attacks) by the Physicians' Health Study Research Group at Harvard Medical School. The Physicians' Health Study was a five-year randomized study testing whether regular intake of aspirin reduces mortality from cardiovascular disease. Every other day, physicians participating in the study took either one aspirin tablet or a placebo. The study was blind—the physicians in the study did not know which type of pill they were taking.

Table 2.3 Cross Classification of Aspirin Use and Myocardial Infarction (MI)

	Myocardial Infarction		
Group	Yes	No	Total
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037

Source: Preliminary Report: Findings from the Aspirin Component of the Ongoing Physicians' Health Study. *N. Engl. J. Med., 318*: 262–264 (1988).

We treat the two rows in Table 2.3 as independent binomial samples. Of the $N_1 = 11{,}034$ physicians taking placebo, 189 suffered myocardial infarction (MI) over the course of the study, a proportion of $p_1 = 189/11{,}034 = .0171$. Of the $N_2 = 11{,}037$ physicians taking aspirin, 104 suffered MI, a proportion of $p_2 = .0094$. The sample difference of proportions is $.0171 - .0094 = .0077$. From (2.2.1), this difference has an estimated standard error of

$$\sqrt{\frac{(.0171)(.9829)}{11{,}034} + \frac{(.0094)(.9906)}{11{,}037}} = 0.0015.$$

A 95% confidence interval for the true difference $\pi_1 - \pi_2$ is $.0077 \pm 1.96(0.0015)$, or $.008 \pm 0.003$, or $(.005, .011)$. Since this interval contains only positive values, we conclude that $\pi_1 - \pi_2 > 0$; that is, $\pi_1 > \pi_2$, so taking aspirin appears to diminish the risk of MI.

2.2.3 Relative Risk

A difference between two proportions of a certain fixed size may have greater importance when both proportions are near 0 or 1 than when they are near the middle of the range. Consider a comparison of two drugs on the proportion of subjects who have adverse reactions when using the drug. The difference between .010 and .001 is the same as the difference between .410 and .401, namely .009. The first difference seems more noteworthy, since ten times as many subjects have adverse reactions with one drug as the other. In such cases, the ratio of proportions is also a useful descriptive measure.

In 2×2 tables, the *relative risk* is the ratio of the "success" probabilities for the two groups,

$$\frac{\pi_1}{\pi_2}. \tag{2.2.3}$$

It can be any nonnegative real number. The proportions .010 and .001 have a relative risk of $.010/.001 = 10.0$, whereas the proportions .410 and .401 have a relative risk of $.410/.401 = 1.02$. A relative risk of 1.00 occurs when $\pi_1 = \pi_2$; that is, when response is independent of group.

Two groups with *sample* proportions of p_1 and p_2 have a sample relative risk of p_1/p_2. Its sampling distribution can be highly skewed unless the sample sizes are quite large, so its confidence interval formula is rather complex (Problem 2.12).

For Table 2.3, the sample relative risk is $p_1/p_2 = .0171/.0094 = 1.82$. The sample proportion of MI cases was 82% higher for the group taking placebo. Using computer software (SAS-PROC FREQ), we find that a 95% confidence interval for the true relative risk is (1.43, 2.30). We can be 95% confident that, after five years, the proportion of MI cases for physicians taking placebo is between 1.43 and 2.30 times the proportion of MI cases for physicians taking aspirin.

The confidence interval for the relative risk indicates that the risk of MI is at least 43% higher for the placebo group. The confidence interval (.005, .011) for

the difference of proportions makes it seem as if the two groups differ by a trivial amount, but the relative risk shows that the difference may have important public health implications. Using the difference of proportions alone to compare two groups can be somewhat misleading when the proportions are both close to zero.

It is sometimes informative to compute also the ratio of "failure" probabilities, $(1-\pi_1)/(1-\pi_2)$. This takes a different value than the ratio of the success probabilities. When one of the two outcomes has small probability, normally one computes the ratio of the probabilities for that outcome.

2.3 THE ODDS RATIO

We next present another measure of association for 2×2 contingency tables, called the *odds ratio*. This is a fundamental parameter for models presented in later chapters.

In 2×2 tables, the probability of "success" is π_1 in row 1 and π_2 in row 2. Within row 1, the *odds* of success are defined to be

$$\text{odds}_1 = \frac{\pi_1}{(1-\pi_1)}.$$

For instance, if $\pi_1 = .75$, then the odds of success equal $.75/.25 = 3$.

The odds are nonnegative, with value greater than 1.0 when a success is more likely than a failure. When odds $= 4.0$, a success is four times as likely as a failure. The probability of success is .8, the probability of failure is .2, and the odds equal $.8/.2 = 4$. We then expect to observe four successes for every one failure. When odds $= \frac{1}{4}$, a failure is four times as likely as a success; we expect to observe one success for every four failures.

Within row 2, the odds of success equal

$$\text{odds}_2 = \frac{\pi_2}{1-\pi_2}.$$

In either row, the success probability is the function of the odds,

$$\pi = \frac{\text{odds}}{\text{odds} + 1}.$$

For instance, when odds $= 4$, then $\pi = 4/(4+1) = .8$. When the conditional distributions are identical in the two rows (i.e., $\pi_1 = \pi_2$), the odds satisfy $\text{odds}_1 = \text{odds}_2$. The variables are then independent.

The ratio of odds from the two rows,

$$\theta = \frac{\text{odds}_1}{\text{odds}_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}, \tag{2.3.1}$$

is called the *odds ratio*. Whereas the relative risk is a ratio of two probabilities, the odds ratio θ is a ratio of two odds.

2.3.1 Properties of the Odds Ratio

The odds ratio can equal any nonnegative number. When X and Y are independent, $\pi_1 = \pi_2$, so that $\text{odds}_1 = \text{odds}_2$ and $\theta = \text{odds}_1/\text{odds}_2 = 1$. The value $\theta = 1$ corresponding to independence serves as a baseline for comparison. Odds ratios on each side of 1 reflect certain types of associations. When $1 < \theta < \infty$, the odds of success are higher in row 1 than in row 2. For instance, when $\theta = 4$, the odds of success in row 1 are four times the odds of success in row 2. Thus, subjects in row 1 are more likely to have successes than are subjects in row 2; that is, $\pi_1 > \pi_2$. When $0 < \theta < 1$, a success is less likely in row 1 than in row 2; that is, $\pi_1 < \pi_2$.

Values of θ farther from 1.0 in a given direction represent stronger levels of association. An odds ratio of 4 is farther from independence than an odds ratio of 2, and an odds ratio of 0.25 is farther from independence than an odds ratio of 0.50. Two values for θ represent the same level of association, but in opposite directions, when one value is the inverse of the other. When $\theta = 0.25$, for instance, the odds of success in row 1 are 0.25 times the odds of success in row 2, or equivalently $1/0.25 = 4.0$ times as high in row 2 as in row 1. When the order of the rows is reversed or the order of the columns is reversed, the new value of θ is the inverse of the original value. This ordering is usually arbitrary, so whether we get 4.0 or 0.25 for the odds ratio is simply a matter of how we label the rows and columns.

The odds ratio does not change value when the orientation of the table reverses so that the rows become the columns and the columns become the rows. The same value occurs when we treat the columns as the response variable and the rows as the explanatory variable, or the rows as the response variable and the columns as the explanatory variable. Since the odds ratio treats the variables symmetrically, it is unnecessary to identify one classification as a response variable in order to calculate it. By contrast, the relative risk requires this, and its value also depends on whether we apply it to the first or second response category.

When both variables are responses, the odds ratio can be defined using joint probabilities as

$$\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}. \tag{2.3.2}$$

The odds ratio is also called the *cross-product ratio*, since it equals the ratio of the products $\pi_{11}\pi_{22}$ and $\pi_{12}\pi_{21}$ of cell probabilities from diagonally opposite cells.

The sample odds ratio equals the ratio of the sample odds in the two rows,

$$\hat{\theta} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}. \tag{2.3.3}$$

For the standard sampling schemes, this is the ML estimator of the true odds ratio.

2.3.2 Odds Ratio for Aspirin Study

To illustrate the odds ratio, we revisit Table 2.3 from Section 2.2.2 on aspirin use and myocardial infarction (MI). For the physicians taking placebo, the estimated

odds of MI equal $n_{11}/n_{12} = 189/10{,}845 = 0.0174$. The value 0.0174 means there were 1.74 "yes" responses for every 100 "no" responses. The estimated odds equal $104/10{,}933 = 0.0095$ for those taking aspirin, or 0.95 "yes" responses per every 100 "no" responses.

The sample odds ratio equals $\hat{\theta} = 0.0174/0.0095 = 1.832$. This also equals the cross-product ratio $(189)(10{,}933)/(10{,}845)(104)$. The estimated odds of MI for physicians taking placebo equal 1.832 times the estimated odds for physicians taking aspirin. The estimated odds were 83% higher for the placebo group.

2.3.3 Inference for Odds Ratios and Log Odds Ratios

For small to moderate sample sizes, the sampling distribution of the odds ratio is highly skewed. When $\theta = 1$, for instance, $\hat{\theta}$ cannot be much smaller than θ (since $\hat{\theta} \geq 0$), but it could be much larger with nonnegligible probability.

Because of this skewness, statistical inference for the odds ratio uses an alternative but equivalent measure: its natural logarithm, $\log(\theta)$. Independence corresponds to $\log(\theta) = 0$. That is, an odds ratio of 1.0 is equivalent to a log odds ratio of 0.0. An odds ratio of 2.0 has a log odds ratio of 0.7. The log odds ratio is symmetric about zero, in the sense that reversal of rows or reversal of columns changes its sign. Two values for $\log(\theta)$ that are the same except for sign, such as $\log(2.0) = 0.7$ and $\log(0.5) = -0.7$, represent the same level of association. Doubling a log odds ratio corresponds to squaring an odds ratio. For instance, log odds ratios of $2(0.7) = 1.4$ and $2(-0.7) = -1.4$ correspond to odds ratios of $2^2 = 4$ and $0.5^2 = 0.25$.

The log transform of the sample odds ratio, $\log \hat{\theta}$, has a less skewed sampling distribution that is closer to normality. Its large-sample approximating normal distribution has a mean of $\log \theta$ and a standard deviation, referred to as an *asymptotic standard error* and denoted by *ASE*, of

$$ASE(\log \hat{\theta}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}. \qquad (2.3.4)$$

The *ASE* value decreases as the cell counts increase. Because this sampling distribution is closer to normality, it is best to construct confidence intervals for $\log \theta$ and then transform back (i.e., take antilogs, using the exponential function) to form a confidence interval for θ. A large-sample confidence interval for $\log \theta$ is

$$\log \hat{\theta} \pm z_{\alpha/2}\, ASE(\log \hat{\theta}).$$

Exponentiating endpoints of this confidence interval yields one for θ.

For Table 2.3, the natural log of $\hat{\theta}$ equals $\log(1.832) = 0.605$. The *ASE* (2.3.4) of $\log \hat{\theta}$ equals $(1/189 + 1/10{,}933 + 1/10{,}845 + 1/104)^{1/2} = 0.123$. For the population this sample represents, a 95% confidence interval for $\log \theta$ equals $0.605 \pm 1.96(0.123)$, or $(0.365, 0.846)$. The corresponding confidence interval for θ is $[\exp(0.365), \exp(0.846)] = (e^{0.365}, e^{0.846}) = (1.44, 2.33)$. Since the confidence interval for θ does not contain 1.0, the true odds of MI seem different for the two

groups. The interval predicts that the odds of MI are at least 44% higher for subjects taking placebo than for subjects taking aspirin. The endpoints of the interval are not equally distant from $\hat{\theta} = 1.83$, because the sampling distribution of $\hat{\theta}$ is skewed to the right.

The sample odds ratio $\hat{\theta}$ equals 0 or ∞ if any $n_{ij} = 0$, and it is undefined if both entries in a row or column are zero. The slightly amended estimator

$$\tilde{\theta} = \frac{(n_{11} + 0.5)(n_{22} + 0.5)}{(n_{12} + 0.5)(n_{21} + 0.5)},$$

corresponding to adding $\frac{1}{2}$ to each cell count, does not have this problem. It is preferred when the cell counts are very small or any zero cell counts occur. In that case, the *ASE* formula (2.3.4) replaces $\{n_{ij}\}$ by $\{n_{ij} + 0.5\}$. For Table 2.3, $\tilde{\theta} = (189.5)(10{,}933.5)/(10{,}845.5)(104.5) = 1.828$ is close to $\hat{\theta} = 1.832$, since no cell count is especially small.

2.3.4 Relationship Between Odds Ratio and Relative Risk

A sample odds ratio of 1.83 does *not* mean that p_1 is 1.83 times p_2; that would be the interpretation of a *relative risk* of 1.83, since that measure deals with proportions rather than odds. Instead, $\hat{\theta} = 1.83$ means that the *odds* value $p_1/(1-p_1)$ is 1.83 times the odds value $p_2/(1-p_2)$. From (2.3.3) and from the sample analog of definition (2.2.3),

$$\text{Odds ratio} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \text{Relative risk} \times \left(\frac{1-p_2}{1-p_1}\right).$$

When the proportion of successes is close to zero for both groups, the fraction in the last term of this expression equals approximately 1.0. The odds ratio and relative risk then take similar values. Table 2.3 illustrates this similarity. For each group, the sample proportion of MI cases is close to zero. Thus, the sample odds ratio of 1.83 is similar to the sample relative risk of 1.82 obtained in Section 2.2.3. In such a case, an odds ratio of 1.83 *does* mean that p_1 is about 1.83 times p_2.

This relationship between the odds ratio and the relative risk is useful. For some data sets calculation of the relative risk is not possible, yet one can calculate the odds ratio and use it to approximate the relative risk. Table 2.4 is an example of this type. These data refer to a study that investigated the relationship between myocardial infarction and smoking. The first column refers to 262 young and middle-aged women (age <69) admitted to 30 coronary care units in northern Italy with acute MI during the period 1983–1988. Each case was matched with two control patients admitted to the same hospitals with other acute disorders. The controls fall in the second column of the table. All subjects were classified according to whether they had ever been smokers. The "yes" group consists of women who were current smokers or ex-smokers, whereas the "no" group consists of women who never were smokers. We refer to this variable as *smoking status*.

Table 2.4 Cross Classification of Smoking Status and Myocardial Infarction (MI)

Ever Smoker	Myocardial Infarction	Controls
Yes	172	173
No	90	346

Source: A. Gramenzi et al., *J. Epidemiol. and Commun. Health, 43:* 214–217 (1989).
Reprinted with permission of BMJ Publishing Group.

We would normally regard MI as a response variable and smoking status as an explanatory variable. In this study, however, the marginal distribution of MI is fixed by the sampling design, there being two controls for each case. The outcome measured for each subject is whether she ever was a smoker. The study, which uses a *retrospective* design to "look into the past," is called a *case-control study*. Such studies are common in health-related applications, for instance, to ensure a sufficiently large sample of subjects having the disease studied.

We might wish to compare ever-smokers with nonsmokers in terms of the proportion who suffered MI. These proportions refer to the conditional distribution of MI, given smoking status. We cannot estimate such proportions for this data set. For instance, about a third of the sample suffered MI. This is because the study matched each MI case with two controls, and it does not make sense to use $\frac{1}{3}$ as an estimate of the probability of MI. We *can* compute proportions in the reverse direction, for the conditional distribution of smoking status, given myocardial infarction status. For women suffering MI, the proportion who ever were smokers was $172/262 = .656$, while it was $173/519 = .333$ for women who had not suffered MI.

When the sampling design is retrospective, one can construct conditional distributions for the explanatory variable, within levels of the fixed response. It is usually not possible to estimate the probability of the response outcome of interest, or to compute the difference of proportions or relative risk for that outcome. Using Table 2.4, for instance, we cannot estimate the difference between nonsmokers and ever smokers in the probability of suffering MI. We can compute the odds ratio, however. This is because the odds ratio takes the same value when it is defined using the conditional distribution of X given Y as it does when defined (as in (2.3.1)) using the distribution of Y given X; that is, it treats the variables symmetrically. The odds ratio is determined by the conditional distributions in *either* direction, and can be calculated even if we have a study design that measures a response on X within each level of Y. In Table 2.4, the sample odds ratio is $[.656/(1-.656)]/[.333/(1-.333)] = (172\times346)/(173\times90) = 3.82$. The estimated odds of ever being a smoker were about 2 for the MI cases (i.e., $.656/.344$) and about $\frac{1}{2}$ for the controls (i.e., $.333/.667$), yielding an odds ratio of about $2/(1/2) = 4$.

We noted that when the probability that $Y = 1$ is small for each value of X, the odds ratio and relative risk take similar values. Even if we can estimate only conditional probabilities of X given Y, if we expect $P(Y = 1 \mid X)$ to be small, then we can use the sample odds ratio to provide a rough indication of the relative risk. For Table 2.4, we cannot estimate the relative risk of MI or the difference of proportions suffering MI. Since the probability of young or middle-aged women suffering MI is probably small regardless of smoking status, however, the odds ratio value of 3.82 is also a rough estimate of the relative risk. We estimate that women who ever smoked were nearly four times as likely to suffer MI as women who never smoked.

In Table 2.4, it makes sense to treat each column, rather than each row, as a binomial sample. Because of the matching that occurs in case-control studies, however, the binomial samples in the two columns are *dependent* rather than independent. Each observation in column 1 is naturally paired with two of the observations in column 2. Chapter 9 presents specialized methods for analyzing dependent binomial samples.

2.3.5 Types of Observational Studies*

By contrast to the study summarized by Table 2.4, imagine a study where we follow a sample of women for the next 20 years, observing the rates of MI for smokers and nonsmokers. Such a sampling design is *prospective*. There are two types of prospective studies. In *cohort studies*, the subjects make their own choice about which group to join (e.g., whether to be a smoker), and we simply observe in future time who suffers MI. In *clinical trials*, we randomly allocate subjects to the two groups of interest, such as in the aspirin study described in Section 2.2.2, again observing in future time who suffers MI. Yet another approach, a *cross-sectional design*, samples women and classifies them simultaneously on the group classification and their current response. As in a case-control study, we can then get the data at once, rather than waiting for future events.

Case-control, cohort, and cross-sectional studies are called *observational studies*. We observe who chooses each group and who has the outcome of interest. By contrast, a clinical trial is an *experimental study*, the investigator having control over which subjects enter each group, for instance, which subjects take aspirin and which take placebo. Clinical trials have fewer potential pitfalls, because of the use of randomization, but observational studies are often more practical for biomedical and social science research.

2.4 CHI-SQUARED TESTS OF INDEPENDENCE

We next show how to test the null hypothesis (H_0) that cell probabilities equal certain fixed values $\{\pi_{ij}\}$. For a sample of size n with cell counts $\{n_{ij}\}$, the values $\{\mu_{ij} = n\pi_{ij}\}$ are called *expected frequencies*. They represent the values of the expectations $\{E(n_{ij})\}$ when H_0 is true.

This notation refers to two-way tables, but similar notions apply to multiway tables or to a set of counts for a single categorical variable. To illustrate, for n flips of a coin, let π denote the probability of a head and $1-\pi$ the probability of a tail on each flip. The null hypothesis that the coin is balanced corresponds to $\pi = 1-\pi = .5$. The

expected frequency of heads equals $\mu = n\pi = n/2$, which also equals the expected frequency of tails. If H_0 is true, we expect to observe about half heads and half tails.

We compare sample cell counts to the expected frequencies to judge whether the data contradict H_0. If H_0 is true for a two-way table, n_{ij} should be close to μ_{ij} in each cell. The larger the differences $\{n_{ij} - \mu_{ij}\}$, the stronger the evidence against H_0. The test statistics used to make such comparisons have large-sample chi-squared distributions.

2.4.1 Pearson Statistic and the Chi-Squared Distribution

The *Pearson chi-squared statistic* for testing H_0 is

$$X^2 = \sum \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}}. \qquad (2.4.1)$$

It was proposed in 1900 by Karl Pearson, the British statistician known also for the Pearson product-moment correlation, among his many contributions. This statistic takes its minimum value of zero when all $n_{ij} = \mu_{ij}$. For a fixed sample size, greater differences between $\{n_{ij}\}$ and $\{\mu_{ij}\}$ produce larger X^2 values and stronger evidence against H_0.

Since larger X^2 values are more contradictory to H_0, the P-value of the test is the null probability that X^2 is at least as large as the observed value. The X^2 statistic has approximately a chi-squared distribution for large sample sizes. It is difficult to specify what "large" means, but $\{\mu_{ij} \geq 5\}$ is sufficient. The P-value is the chi-squared right-hand tail probability above the observed X^2 value.

The chi-squared distribution is specified by its *degrees of freedom*, denoted by df. The mean of the chi-squared distribution equals df, and its standard deviation equals $\sqrt{2\,df}$. As df increases, the distribution concentrates around larger values and is more spread out. It is defined only for nonnegative values and is skewed to the right, but becomes more bell-shaped (normal) as df increases. Figure 2.1 displays the shapes of chi-squared densities having $df = 1$, 5, 10, and 20. The df value equals the difference between the number of parameters in the alternative and null hypotheses, as explained later in this section.

2.4.2 Likelihood-Ratio Statistic

An alternative statistic for testing H_0 results from the likelihood-ratio method for significance tests. The test determines the parameter values that maximize the likelihood function under the assumption that H_0 is true. It also determines the values that maximize it under the more general condition that H_0 may or may not be true. The test is based on the ratio of the maximized likelihoods,

$$\Lambda = \frac{\text{maximum likelihood when parameters satisfy } H_0}{\text{maximum likelihood when parameters are unrestricted}}.$$

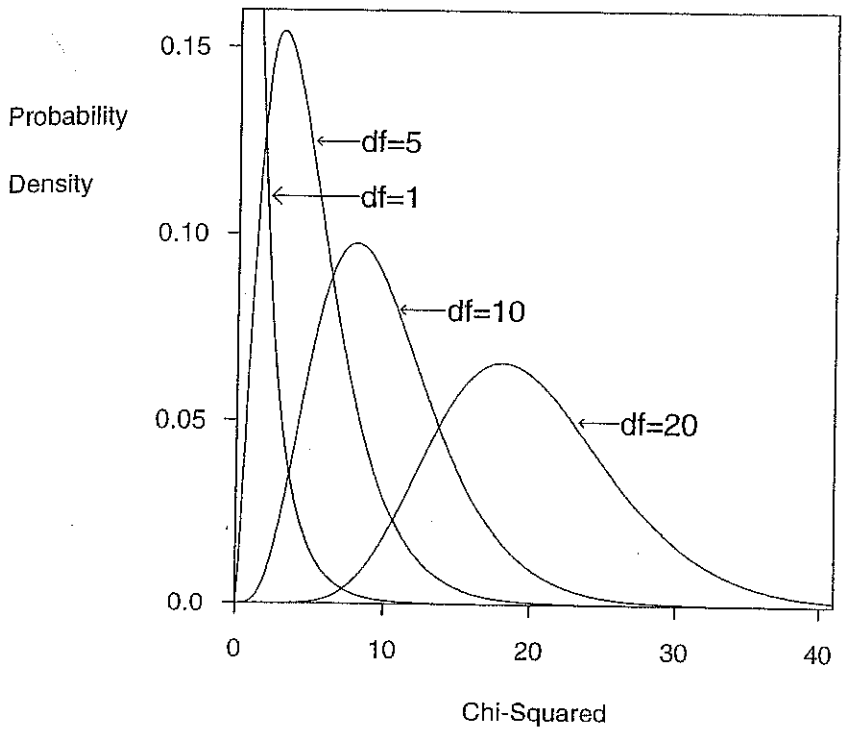

Figure 2.1 Examples of chi-squared distributions.

This ratio cannot exceed 1. If the maximized likelihood is much larger when the parameters are not forced to satisfy H_0, then the ratio Λ is far below 1 and there is strong evidence against H_0.

The test statistic for a likelihood-ratio test equals $-2\log(\Lambda)$. This value is nonnegative, and "small" values of Λ yield "large" values of $-2\log(\Lambda)$. The reason for the log transform is to yield an approximate chi-squared sampling distribution. For two-way contingency tables, this statistic simplifies to the formula

$$G^2 = 2\sum n_{ij} \log\left(\frac{n_{ij}}{\mu_{ij}}\right). \qquad (2.4.2)$$

The statistic G^2 is called the *likelihood-ratio chi-squared statistic*. Like the Pearson statistic, G^2 takes its minimum value of 0 when all $n_{ij} = \mu_{ij}$, and larger values provide stronger evidence against H_0.

Though the Pearson X^2 and likelihood-ratio G^2 provide separate test statistics, they share many properties and commonly yield the same conclusions. When H_0 is

true and the sample cell counts are large, the two statistics have the same chi-squared distribution, and their numerical values are similar. Each statistic has advantages and disadvantages, which we allude to later in this section and in Sections 7.3.1 and 7.4.3.

2.4.3 Tests of Independence

In two-way contingency tables, the null hypothesis of statistical independence of two responses has the form

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$$

for all i and j. The marginal probabilities then specify the joint probabilities. To test H_0, we identify $\mu_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j}$ as the expected frequency. Here, μ_{ij} is the expected value of n_{ij} assuming independence. Usually, $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$ are unknown, as is this expected value.

We estimate the expected frequencies by substituting sample proportions for the unknown probabilities, giving

$$\hat{\mu}_{ij} = np_{i+}p_{+j} = n\frac{n_{i+}}{n}\frac{n_{+j}}{n} = \frac{n_{i+}n_{+j}}{n}.$$

The $\{\hat{\mu}_{ij}\}$ are called *estimated expected frequencies*. They have the same row and column totals as the observed counts, but they display the pattern of independence.

For testing independence in $I \times J$ contingency tables, the Pearson and likelihood-ratio statistics equal

$$X^2 = \sum \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}, \qquad G^2 = 2\sum n_{ij}\log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right). \tag{2.4.3}$$

Their large-sample chi-squared distributions have $df = (I-1)(J-1)$. This means the following: Under H_0, $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$ determine the cell probabilities. There are $I-1$ nonredundant row probabilities; since they sum to 1, the first $I-1$ determine the last one through $\pi_{I+} = 1 - (\pi_{1+} + \cdots + \pi_{I-1,+})$. Similarly, there are $J-1$ nonredundant column probabilities, for a total of $(I-1)+(J-1)$ parameters. The alternative hypothesis does not specify the IJ cell probabilities. They are then solely constrained to sum to 1, so there are $IJ-1$ nonredundant parameters. The value for df is the difference between the number of parameters under the alternative and null hypotheses, or

$$(IJ-1) - [(I-1)+(J-1)] = IJ - I - J + 1 = (I-1)(J-1).$$

2.4.4 Gender Gap Example

We illustrate chi-squared tests of independence using Table 2.5, from the 1991 General Social Survey. The variables are gender and party identification. Subjects indicated whether they identified more strongly with the Democratic or Republican party

Table 2.5 Cross Classification of Party Identification by Gender

	Party Identification			
Gender	Democrat	Independent	Republican	Total
Females	279	73	225	577
	(261.4)	(70.7)	(244.9)	
Males	165	47	191	403
	(182.6)	(49.3)	(171.1)	
Total	444	120	416	980

Note: Estimated expected frequencies for hypothesis of independence in parentheses.
Source: Data from 1991 General Social Survey.

or as Independents. Table 2.5 also contains estimated expected frequencies for H_0: independence. For instance, the first cell has $\hat{\mu}_{11} = n_{1+}n_{+1}/n = (577\times444)/980 = 261.4$.

The chi-squared test statistics are $X^2 = 7.01$ and $G^2 = 7.00$, based on $df = (I-1)(J-1) = (2-1)(3-1) = 2$. The reference chi-squared distribution has a mean of $df = 2$ and a standard deviation of $\sqrt{2\,df} = \sqrt{4} = 2$, so a value of 7.0 is fairly far out in the right-hand tail. Each statistic has a P-value of .03. This evidence of association would be rather unusual if the variables were truly independent. Both test statistics suggest that party identification and gender are associated.

Most major statistical software packages have routines for calculating X^2, G^2, and their P-values. These P-values are approximations for true P-values, since the chi-squared distribution is an approximation for the true sampling distribution. Thus, it would be overly optimistic for us to report P-values to the 4 or 5 decimal places that software provides them. If we are lucky, the P-value approximation is good to the second decimal place, so it makes more sense to report it as .03 (or, at best, .028) rather than .02837. In any case, a P-value simply summarizes the strength of evidence against the null hypothesis, and accuracy to two or three decimal places is sufficient for this purpose.

2.4.5 Residuals

A test statistic and its P-value simply describe the evidence against the null hypothesis. A cell-by-cell comparison of observed and estimated expected frequencies helps us better understand the nature of the evidence. Larger differences between n_{ij} and $\hat{\mu}_{ij}$ tend to occur for cells that have larger expected frequencies, so the raw difference $n_{ij} - \hat{\mu}_{ij}$ is insufficient. For the test of independence, useful cell residuals have the form

$$\frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1-p_{i+})(1-p_{+j})}} \tag{2.4.4}$$

These are called *adjusted residuals*.

When the null hypothesis is true, each adjusted residual has a large-sample standard normal distribution. An adjusted residual that exceeds about 2 or 3 in absolute value

Table 2.6 Adjusted Residuals (in Parentheses) for Testing Independence in Table 2.5

	Party Identification		
Gender	Democrat	Independent	Republican
Females	279	73	225
	(2.29)	(0.46)	(−2.62)
Males	165	47	191
	(−2.29)	(−0.46)	(2.62)

indicates lack of fit of H_0 in that cell. Table 2.6 shows the adjusted residuals for testing independence in Table 2.5. For the first cell, for instance, $n_{11} = 279$ and $\hat{\mu}_{11} = 261.4$. The first row and first column marginal proportions equal $p_{1+} = 577/980 = .589$ and $p_{+1} = 444/980 = .453$. Substituting into (2.4.4), the adjusted residual for this cell equals

$$\frac{279 - 261.4}{\sqrt{261.4(1 - .589)(1 - .453)}} = 2.29.$$

This cell shows a greater discrepancy between n_{11} and $\hat{\mu}_{11}$ than one would expect if the variables were truly independent.

Table 2.6 shows large positive residuals for female Democrats and male Republicans, and large negative residuals for female Republicans and male Democrats. Thus, there were significantly more female Democrats and male Republicans and fewer female Republicans and male Democrats than the hypothesis of independence predicts. An odds ratio describes this evidence of a gender gap. The 2×2 table of Democrat and Republican identifiers has a sample odds ratio of $(279)(191)/(225)(165) = 1.44$. Of those subjects identifying with one of the two parties, the estimated odds of identifying with the Democrats rather than the Republicans were 44% higher for females than males.

For each party, Table 2.6 shows that there is only one nonredundant residual; the one for females is the negative of the one for males. The observed counts and the estimated expected frequencies have the same row and column totals. Thus, in a given column, if $n_{ij} > \hat{\mu}_{ij}$ in one cell, the reverse must happen in the other cell. The differences $n_{1j} - \hat{\mu}_{1j}$ and $n_{2j} - \hat{\mu}_{2j}$ have the same magnitude but different signs, implying the same pattern for their adjusted residuals.

2.4.6 Partitioning Chi-Squared

Chi-squared statistics have a reproductive property. If one chi-squared statistic has $df = df_1$ and a separate, independent, chi-squared statistic has $df = df_2$, then their sum is chi-squared with $df = df_1 + df_2$. For instance, if we had a table of form Table 2.5 for college-educated subjects and a separate one for subjects not having a college education, the sum of the X^2 values or the sum of the G^2 values from the two tables would be a chi-squared statistic with $df = 2 + 2 = 4$.

Similarly, chi-squared statistics having $df > 1$ can be broken into components with fewer degrees of freedom. For instance, a statistic having $df = 2$ can be partitioned into two independent components each having $df = 1$. Another supplement to a test of independence partitions its chi-squared test statistic so that the components represent certain aspects of the association. A partitioning may show that an association primarily reflects differences between certain categories or groupings of categories.

We illustrate with a partitioning of G^2 for testing independence in $2 \times J$ tables. The test statistic then has $df = (J - 1)$, and we partition it into $J - 1$ components. The jth component is G^2 for testing independence in a 2×2 table, where the first column combines columns 1 through j of the original table, and the second column uses column $j + 1$ of the original table. That is, G^2 for testing independence in a $2 \times J$ table equals the sum of a G^2 statistic that compares the first two columns, plus a G^2 statistic for the 2×2 table that combines the first two columns and compares them to the third column, and so on, up to a G^2 statistic for the 2×2 table that combines the first $J - 1$ columns and compares them to the last column. Each component G^2 statistic has $df = 1$.

Consider again Table 2.5. The first two columns of this table form a 2×2 table with cell counts, by row, of $(279, 73/165, 47)$. For this component table, $G^2 = 0.16$, with $df = 1$. Of those subjects who identify either as Democrats or Independents, there is little evidence of a difference between females and males in the relative numbers in the two categories. We form the second 2×2 table by combining these columns and comparing them to the Republican column, giving the table with rows $(279 + 73, 225/165 + 47, 191) = (352, 225/212, 191)$. This table has $G^2 = 6.84$, based on $df = 1$. There is strong evidence of a difference between females and males in the relative numbers identifying as Republican instead of Democrat or Independent. Note that $0.16 + 6.84 = 7.00$; that is, the sum of these G^2 components equals G^2 for the test of independence for the complete 2×3 table. This overall statistic primarily reflects differences between genders in choosing between Republicans and Democrats/Independents.

It might seem more natural to compute G^2 for separate 2×2 tables that pair each column with a particular one, say the last. Though this is a reasonable way to investigate association in many data sets, these component statistics are not independent and do not sum to G^2 for the complete table. Certain rules determine ways of forming tables so that chi-squared partitions, but they are beyond the scope of this text (see, e.g., Agresti (1990), p. 53, for rules and references). A necessary condition is that the G^2 values for the component tables sum to G^2 for the original table.

The G^2 statistic has exact partitionings. The overall Pearson X^2 statistic does not equal the sum of the X^2 values for the separate tables in a partition. However, it is valid to use the X^2 statistics for the separate tables in the partition; they simply do not provide an exact algebraic partitioning of the X^2 statistic for the overall table.

2.4.7 Comments on Chi-Squared Tests

Chi-squared tests of independence, like any significance tests, have serious limitations. They simply indicate the degree of evidence for an association. They are rarely

adequate for answering all questions we have about a data set. Rather than relying solely on results of these tests, one should study the nature of the association. It is sensible to decompose chi-squared into components, study residuals, and estimate parameters such as odds ratios that describe the strength of association.

The X^2 and G^2 chi-squared tests also have limitations in the types of data sets for which they are applicable. For instance, they require large samples. The sampling distributions of X^2 and G^2 get closer to chi-squared as the sample size n increases, relative to the number of cells IJ. The convergence is quicker for X^2 than G^2. The chi-squared approximation is often poor for G^2 when $n/IJ < 5$. When I or J is large, it can be decent for X^2 when some expected frequencies are as small as 1. Section 7.4.3 provides further guidelines, but these are not crucial since small-sample procedures are available whenever we question whether n is sufficiently large. Section 2.6 discusses these.

The $\{\hat{\mu}_{ij} = n_{i+}n_{+j}/n\}$ used in X^2 and G^2 depend on the row and column marginal totals, but not on the order in which the rows and columns are listed. Thus, X^2 and G^2 do not change value with arbitrary reorderings of rows or of columns. This means that these tests treat both classifications as nominal. We ignore some information when we use them to test independence between ordinal classifications. When at least one variable is ordinal, more powerful tests of independence usually exist. The next section presents such a test.

2.5 TESTING INDEPENDENCE FOR ORDINAL DATA

The chi-squared test of independence using test statistic X^2 or G^2 treats both classifications as nominal. When the rows and/or the columns are ordinal, test statistics that utilize the ordinality are usually more appropriate.

2.5.1 Linear Trend Alternative to Independence

When the row variable X and the column variable Y are ordinal, a "trend" association is quite common. As the level of X increases, responses on Y tend to increase toward higher levels, or responses on Y tend to decrease toward lower levels. One can use a single parameter to describe such an ordinal trend association. The most common analysis assigns scores to categories and measures the degree of *linear trend* or correlation.

We next present a test statistic that is sensitive to positive or negative linear trends in the relationship between X and Y. It utilizes correlation information in the data. Let $u_1 \leq u_2 \leq \cdots \leq u_I$ denote scores for the rows, and let $v_1 \leq v_2 \leq \cdots \leq v_J$ denote scores for the columns. The scores have the same ordering as the category levels and are said to be *monotone*. The scores reflect distances between categories, with greater distances between categories treated as farther apart.

The sum $\sum_{i,j} u_i v_j n_{ij}$, which weights cross-products of scores by the frequency of their occurrence, relates to the covariation of X and Y. For the chosen scores, the Pearson product-moment correlation between X and Y equals the standardization of this sum,

$$r = \frac{\sum_{i,j} u_i v_j n_{ij} - \left(\sum_i u_i n_{i+}\right)\left(\sum_j v_j n_{+j}\right)/n}{\sqrt{\left[\sum_i u_i^2 n_{i+} - \frac{\left(\sum_i u_i n_{i+}\right)^2}{n}\right]\left[\sum_j v_j^2 n_{+j} - \frac{\left(\sum_j v_j n_{+j}\right)^2}{n}\right]}}.$$

Alternative formulas exist for r, and one can compute it using standard software, entering for each subject their score on the row classification and their score on the column classification. The correlation falls between -1 and $+1$. Independence between the variables implies that its true value equals zero. The larger the correlation is in absolute value, the farther the data fall from independence in this linear dimension.

A statistic for testing the null hypothesis of independence against the two-sided alternative hypothesis of nonzero true correlation is given by

$$M^2 = (n-1)r^2. \tag{2.5.1}$$

This statistic increases as the sample correlation r increases in magnitude and as the sample size n grows. For large samples, it has approximately a chi-squared distribution with $df = 1$. Large values contradict independence, so, as with X^2 and G^2, the P-value is the right-tail probability above the observed value. The square root, $M = \sqrt{n-1}r$, has approximately a standard normal null distribution. It applies to directional alternatives, such as positive correlation between the classifications.

Tests using M^2 treat the variables symmetrically. If one interchanges the rows with the columns and their scores in an $I \times J$ table, M^2 takes identical value for the corresponding $J \times I$ table.

2.5.2 Alcohol and Infant Malformation Example

Table 2.7 refers to a prospective study of maternal drinking and congenital malformations. After the first three months of pregnancy, the women in the sample completed a questionnaire about alcohol consumption. Following childbirth, observations were recorded on presence or absence of congenital sex organ malformations. Alcohol

Table 2.7 Infant Malformation and Mother's Alcohol Consumption

Alcohol Consumption	Malformation		Total	Percentage Present	Adjusted Residual
	Absent	Present			
0	17,066	48	17,114	0.28	−0.18
< 1	14,464	38	14,502	0.26	−0.71
1–2	788	5	793	0.63	1.84
3–5	126	1	127	0.79	1.06
≥ 6	37	1	38	2.63	2.71

Source: B. I. Graubard and E. L. Korn, *Biometrics 43:* 471–476 (1987).
Reprinted with permission of the Biometric Society.

consumption, measured as average number of drinks per day, is an explanatory variable with ordered categories. Malformation, the response variable, is nominal. When a variable is nominal but has only two categories, statistics (such as M^2) that treat the variable as ordinal are still valid. For instance, we could artificially regard malformation as ordinal, treating "absent" as "low" and "present" as "high." Any choice of two scores yields the same value of M^2, and we simply use 0 for "absent" and 1 for "present."

Table 2.7 has a mixture of very small, moderate, and extremely large counts. Even though the sample size is large ($n = 32{,}574$), in such cases the actual sampling distributions of X^2 or G^2 may not be close to chi-squared. For these data, having $df = 4$, $G^2 = 6.2$ ($P = .19$) and $X^2 = 12.1$ ($P = .02$), so they provide mixed signals. In any case, they ignore the ordinality of alcohol consumption.

Table 2.7 lists the percentage of malformation cases at each level of alcohol consumption. These percentages show roughly an increasing trend. The first two are similar and the next two are also similar, however, and any of the last three percentages changes dramatically with the addition or deletion of one malformation case. Table 2.7 also reports adjusted residuals for the "present" category in this table. They are negative at low levels of alcohol consumption and positive at high levels of consumption, though most are small, and they also change substantially with slight changes in the data. The sample percentages and the adjusted residuals both suggest a possible tendency for malformations to be more likely at higher levels of alcohol consumption.

The ordinal test statistic M^2 requires scores for levels of alcohol consumption. It seems sensible to use scores that are midpoints of the categories; that is, $v_1 = 0$, $v_2 = 0.5$, $v_3 = 1.5$, $v_4 = 4.0$, $v_5 = 7.0$, the last score being somewhat arbitrary. One can calculate r and M^2 using software (e.g., PROC FREQ in SAS; see Table A.2 in the Appendix). The sample correlation between alcohol consumption and malformation is $r = .014$, and $M^2 = (32{,}573)(.014)^2 = 6.6$. The P-value of .01 suggests strong evidence of a nonzero correlation. The standard normal statistic $M = 2.56$ has $P = .005$ for the one-sided alternative of a positive correlation.

For the chosen scores, the correlation value of .014 seems weak. However, r has limited use as a descriptive measure for tables, such as this one, that are highly discrete and unbalanced. Future chapters present tests such as M^2 as part of a model-based analysis. For instance, Section 4.2 presents a model in which the probability of malformation changes linearly according to alcohol consumption. Model-based approaches yield estimates of the size of the effect as well as smoothed estimates of cell probabilities. These estimates are more informative than mere significance tests.

2.5.3 Extra Power with Ordinal Test

For testing independence, X^2 and G^2 refer to the most general alternative hypothesis possible, whereby cell probabilities exhibit *any* type of statistical dependence. Their df value of $(I-1)(J-1)$ reflects an alternative hypothesis that has $(I-1)(J-1)$ more parameters than the null hypothesis. These statistics are designed to detect any type of pattern for the additional parameters. In achieving this generality, they sacrifice sensitivity for detecting particular patterns.

When the row and column variables are ordinal, one can attempt to describe the association using a single extra parameter. For instance, the test statistic M^2 is based on a correlation measure of linear trend. When a test statistic refers to a single parameter, it has $df = 1$.

When the association truly has a positive or negative trend, the ordinal test using M^2 has a power advantage over the tests based on X^2 or G^2. Since df equals the mean of the chi-squared distribution, a relatively large M^2 value based on $df = 1$ falls farther out in its right-hand tail than a comparable value of X^2 or G^2 based on $df = (I-1)(J-1)$; falling farther out in the tail produces a smaller P-value. When there truly is a linear trend, M^2 tends to have similar size as X^2 or G^2, so it tends to have greater power in terms of yielding smaller P-values. In attempting to detect any type of dependence, the X^2 and G^2 statistics lose power relative to statistics designed to detect a particular type of dependence if that type of dependence truly occurs.

Another advantage of chi-squared tests having small df values relates to the accuracy of chi-squared approximations. For small to moderate sample sizes, the true sampling distributions tend to be closer to chi-squared when df is smaller. When several cell counts are small, the chi-squared approximation is likely to be worse for X^2 or G^2 than it is for M^2.

2.5.4 Choice of Scores

For most data sets, the choice of scores has little effect on the results. Different choices of monotone scores usually give similar results. This may not happen, however, when the data are very unbalanced, such as when some categories have many more observations than other categories. Table 2.7 illustrates this. For the equally-spaced row scores (1, 2, 3, 4, 5), the test statistic equals $M^2 = 1.83$, giving a much weaker conclusion ($P = .18$). The magnitudes of r and M^2 do not change with transformations of the scores that maintain the same relative spacings between the categories. For instance, scores $(1, 2, 3, 4, 5)$ yield the same correlation as scores $(0, 1, 2, 3, 4)$ or $(2, 4, 6, 8, 10)$ or $(10, 20, 30, 40, 50)$.

An alternative approach avoids the responsibility of selecting scores and uses the data to form them automatically. Specifically, one assigns ranks to the subjects and uses them as the category scores. For all subjects in a category, one assigns the average of the ranks that would apply for a complete ranking of the sample from 1 to n. These are called *midranks*. We illustrate by assigning midranks to the levels of alcohol consumption in Table 2.7. The 17,114 subjects at level 0 for alcohol consumption share ranks 1 through 17,114. We assign to each of them the average of these ranks, which is the midrank $(1 + 17{,}114)/2 = 8557.5$. The 14,502 subjects at level < 1 for alcohol consumption share ranks 17,115 through $17{,}114 + 14{,}502 = 31{,}616$, for a midrank of $(17{,}115 + 31{,}616)/2 = 24{,}365.5$. Similarly the midranks for the last three categories are 32,013.0, 32,473.0, and 32,555.5. These scores yield $M^2 = 0.35$ and a weaker conclusion yet: ($P = .55$).

Why does this happen? Adjacent categories having relatively few observations necessarily have similar midranks. For instance, the midranks (8557.5, 24,365.5, 32,013.0, 32,473.0, 32,555.5) for Table 2.7 are similar for the final three categories, since those categories have considerably fewer observations than the first two

categories. A consequence is that this scoring scheme treats alcohol consumption level 1–2 (category 3) as much closer to consumption level ≥ 6 (category 5) than to consumption level 0 (category 1). This seems inappropriate. It is usually better to use one's judgment by selecting scores that reflect distances between categories. When uncertain about this choice, perform a sensitivity analysis. Select two or three "sensible" choices and check that the results are similar for each. Equally-spaced scores often provide a reasonable compromise when the category labels do not suggest any obvious choices, such as the categories (liberal, moderate, conservative) for political philosophy.

When X and Y are both ordinal, one can use midrank scores for each. The M^2 statistic is then sensitive to detecting nonzero values of a nonparametric form of correlation called *Spearman's rho*. Alternative ordinal tests for $I \times J$ tables utilize versions of other ordinal association measures. For instance, *gamma* and *Kendall's tau-b* are contingency table generalizations of the ordinal measure called *Kendall's tau*. The sample value of any such measure divided by its standard error has a large-sample standard normal distribution for testing independence, and the square of the statistic is chi-squared with $df = 1$. Like the test based on M^2, these tests share the potential power advantage that results from using a single parameter to describe the association.

2.5.5 Trend Tests for I-by-2 and 2-by-J Tables

We now study how M^2 utilizes the sample data when X or Y has only two levels. Suppose the row variable X is an explanatory variable, and the column variable Y is a response variable.

When X is binary, the table has size $2 \times J$. Tables of this size occur in comparisons of two groups, such as when the rows represent two treatments. Using scores ($u_1 = 0$, $u_2 = 1$) for levels of X in this case, we see that the covariation measure $\sum_{i,j} u_i v_j n_{ij}$ on which M^2 is based simplifies to $\sum_j v_j n_{2j}$. This term sums the scores on Y for all subjects in row 2. Divided by the number of subjects in row 2, it gives the mean score for that row. In fact, when the columns (Y) are ordinal with scores $\{v_j\}$, the M^2 statistic for $2 \times J$ tables is directed toward detecting differences between the two row means of the scores on Y. In testing independence using M^2, small P-values suggest that the true difference in row means is nonzero.

When we use midrank scores for Y, the test for $2 \times J$ tables is sensitive to differences in mean ranks for the two rows. This test is called the *Wilcoxon* or *Mann–Whitney test*. Most nonparametric statistics texts present this test for fully-ranked response data, whereas the $2 \times J$ table is an extended case in which sets of subjects at the same level of Y are tied and use midranks. The large-sample version of that nonparametric test uses a standard normal z statistic. The square of the z statistic is equivalent to M^2, using arbitrary scores (such as 0, 1) for the rows and midranks for the columns.

Tables of size $I \times 2$, such as Table 2.7, have a binary response variable rather than a binary explanatory variable. It is then natural to focus on how the proportion classified in a given response category of Y varies across the levels of X. For ordinal X with monotone row scores and arbitrary scores for the two columns, M^2 focuses on detecting a linear trend in this proportion and relates to models presented in Section 4.2. In testing independence using M^2, small P-values suggest that the slope for this linear trend is nonzero. This $I \times 2$ version of the ordinal test is called the *Cochran-Armitage trend test*.

2.5.6 Nominal-Ordinal Tables

The test statistic (2.5.1) treats both classifications as ordinal. When one variable (say X) is nominal but has only two categories, we can still use it. When X is nominal with more than two categories, it is inappropriate, and we use a different statistic. It is based on calculating a mean response on the ordinal variable in each row and considering the variation among the row means. The statistic is rather complex computationally, and we defer discussion of it to Section 7.3.6. It has a large-sample chi-squared distribution with $df = (I - 1)$. When $I = 2$, it is identical to M^2, which then compares the two row means.

2.6 EXACT INFERENCE FOR SMALL SAMPLES

The confidence intervals and tests presented so far in this chapter are large-sample methods. As the sample size n grows, the cell counts grow, and "chi-squared" statistics such as X^2, G^2, and M^2 have distributions that are more nearly chi-squared. When the sample size is small, one can perform inference using *exact* distributions rather than large-sample approximations. This section discusses exact inference for two-way contingency tables.

2.6.1 Fisher's Exact Test

We first study the 2×2 case. The null hypothesis of independence corresponds to an odds ratio of $\theta = 1$. A small-sample probability distribution for the cell counts is defined for the set of tables having the same row and column totals as the observed data. Under Poisson, binomial, or multinomial sampling assumptions for the cell counts, the distribution that applies to this restricted set of tables fixing the row and column totals is called the *hypergeometric*.

For given row and column marginal totals, the value for n_{11} determines the other three cell counts. Thus, the hypergeometric formula expresses probabilities for the four cell counts in terms of n_{11} alone. When $\theta = 1$, the probability of a particular value n_{11} for that count equals

$$P(n_{11}) = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{+1} - n_{11}}}{\binom{n}{n_{+1}}}. \qquad (2.6.1)$$

The binomial coefficients equal

$$\binom{a}{b} = \frac{a!}{b!\,(a-b)!}.$$

To test independence, the P-value is the sum of hypergeometric probabilities for outcomes at least as favorable to the alternative hypothesis as the observed outcome. We illustrate for $H_a : \theta > 1$. Given the marginal totals, tables having larger n_{11} values also have larger sample odds ratios $\hat{\theta} = (n_{11}n_{22})/(n_{12}n_{21})$, and hence provide stronger evidence in favor of this alternative. The P-value equals the right-tail hypergeometric probability that n_{11} is at least as large as the observed value. This test for 2×2 tables, proposed by the eminent British statistician R. A. Fisher in 1934, is called *Fisher's exact test.*

2.6.2 Fisher's Tea Taster

To illustrate this test in his 1935 text, *The Design of Experiments*, Fisher described the following experiment: A colleague of Fisher's at Rothamsted Experiment Station near London claimed that, when drinking tea, she could distinguish whether milk or tea was added to the cup first. To test her claim, Fisher designed an experiment in which she tasted eight cups of tea. Four cups had milk added first, and the other four had tea added first. She was told there were four cups of each type, so that she should try to select the four that had milk added first. The cups were presented to her in random order.

Table 2.8 shows a potential result of the experiment. We conduct Fisher's exact test of $H_0 : \theta = 1$ against $H_a : \theta > 1$. The null hypothesis states that Fisher's colleague's guess was independent of the actual order of pouring; the alternative hypothesis reflects her claim, predicting a positive association between true order of pouring and her guess. For this experimental design, the column margins are identical to the row margins (4, 4), since she knew that four cups had milk added first. Both marginal distributions are naturally fixed.

The null distribution of n_{11} is the hypergeometric distribution defined for all 2×2 tables having row and column margins (4, 4). The potential values for n_{11} are (0, 1, 2, 3, 4). The observed table, three correct guesses of the four cups having milk

Table 2.8 Fisher's Tea-Tasting Experiment

	Guess Poured First		
Poured First	Milk	Tea	Total
Milk	3	1	4
Tea	1	3	4
Total	4	4	

added first, has null probability

$$P(3) = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} = \frac{[4!/(3!)(1!)][4!/(1!)(3!)]}{[8!/(4!)(4!)]} = \frac{16}{70} = .229.$$

The only table that is more extreme, for the alternative $H_a : \theta > 1$, consists of four correct guesses. It has $n_{11} = n_{22} = 4$ and $n_{12} = n_{21} = 0$, and a probability of

$$P(4) = \frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = \frac{1}{70} = .014.$$

Table 2.9 summarizes the possible values of n_{11} and their probabilities.

The P-value for the one-sided alternative $H_a : \theta > 1$ equals the right-tail probability that n_{11} is at least as large as observed; that is, $P = P(3) + P(4) = .243$. This is not much evidence against the null hypothesis of independence. The experiment did not establish an association between the actual order of pouring and the guess. Of course, it is difficult to show effects with such a small sample. If the tea taster had guessed all cups correctly (i.e., $n_{11} = 4$), the observed result would have been the most extreme possible in the right-hand tail of the hypergeometric distribution; then, $P = P(4) = .014$, giving some reason to believe her claim. For the potential n_{11} values, Table 2.9 shows P-values for the alternative $H_a : \theta > 1$.

2.6.3 P-values and Type I Error Probabilities

The two-sided alternative $H_a : \theta \neq 1$ refers to the general alternative of statistical dependence used in chi-squared tests. Its exact P-value is usually defined as the two-tailed sum of the probabilities of tables no more likely than the observed table. To calculate it, one adds the hypergeometric probabilities of all outcomes y for the first cell count for which $P(y) \leq P(n_{11})$, where n_{11} is the observed count. For Table 2.8, summing all probabilities that are no greater than the probability $P(3) = .229$ of the

Table 2.9 Hypergeometric Distribution for Tables with Margins of Table 2.8

n_{11}	Probability	P-value	X^2
0	.014	1.000	8.0
1	.229	.986	2.0
2	.514	.757	0.0
3	.229	.243	2.0
4	.014	.014	8.0

Note: P-value refers to right-tail probability for one-sided alternative.

observed table gives $P = P(0) + P(1) + P(3) + P(4) = .486$. When the row or column marginal totals are equal, the hypergeometric distribution is symmetric, and the two-sided P-value doubles the one-sided one.

An alternative two-sided P-value sums the probabilities of those tables for which the Pearson X^2 statistic is at least as large as the observed value. That is, it uses the exact small-sample distribution of X^2 rather than its large-sample chi-squared distribution. Table 2.9 shows the X^2 values for the five tables having the margins of Table 2.8. The statistic can assume only three distinct values, so its highly discrete distribution is far from the continuous chi-squared distribution. Figure 2.2 plots this exact small-sample distribution of X^2. It equals 0.0 with probability .514, 2.0 with probability .458, and 8.0 with probability .028. The observed table has $X^2 = 2.0$, and the P-value equals the null probability of a value this large or larger, or $.458 + .028 = .486$. For these data, this P-value based on X^2 is identical to the one based solely on probabilities.

Computations for the hypergeometric distribution are rather messy. One can sidestep this distribution and approximate the exact P-value for X^2 by obtaining

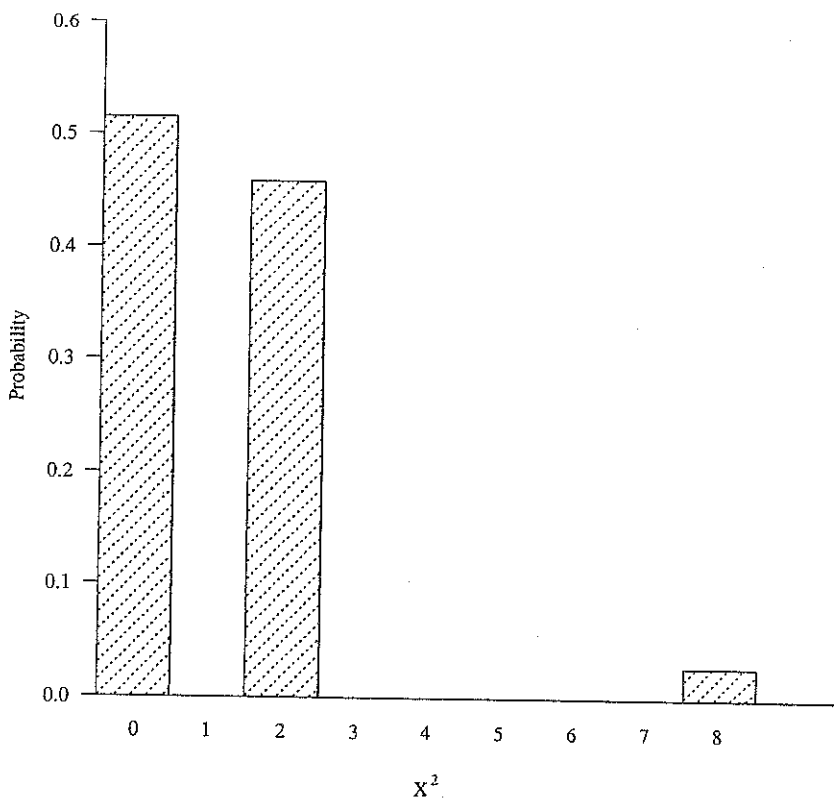

Figure 2.2 Exact distribution of Pearson X^2 for Table 2.8.

a P-value from the chi-squared distribution for an adjustment of the Pearson statistic using the *Yates continuity correction*. There is no longer any reason to use this approximation, however, since modern software makes it possible to conduct Fisher's exact test even for fairly large samples with hypergeometric P-values based on the X^2 or probability criteria.

For small samples, the exact distribution (2.6.1) is highly discrete, in the sense that n_{11} can assume relatively few values. The P-value also has a small number of possible values. For Table 2.8, it can assume five values for the one-sided test and three values for the two-sided test. This has an impact on error rates in hypothesis testing. Suppose we make a formal decision about the null hypothesis using a supposed Type I error probability such as .05. That is, we reject the null hypothesis if the P-value is less than or equal to .05. Because of the test's discreteness, it is usually not possible to achieve that level exactly. For the one-sided alternative, the tea-tasting experiment yields a P-value below .05 only when $n_{11} = 4$, in which case $P = .014$. When H_0 is true, the probability of this outcome is .014, so the actual Type I error probability would be .014, not .05. The test is said to be *conservative*, since the actual error rate is smaller than the intended one. (The approximation of exact tests using the Yates continuity correction is also conservative.)

This illustrates an awkwardness with formal decision-making at "sacred" levels such as .05 when the test statistic is discrete. For test statistics having a *continuous* distribution, the P-value has a *uniform* null distribution over the interval [0, 1]. That is, P is equally likely to fall anywhere between 0 and 1, so the probability that P falls below a fixed level α equals α, and the expected value of P is .5. For test statistics having discrete distributions, the null distribution of the P-value is discrete and has expected value greater than .5. For instance, for the one-sided test with the tea-tasting data, the P-value equals .014 with probability $P(0) = .014$, it equals .243 with probability $P(1) = .229$, and so forth; from Table 2.9, the expected value of the P-value is

$$\sum P \times \text{Prob}(P) = .014(.014) + .243(.229) + .757(.514) + .986(.229) + 1.0(.014)$$
$$= .685.$$

In this average sense, P-values for discrete distributions tend to be too large.

To diminish the conservativeness of tests for discrete data, one can use a slightly different definition of P-value. The *mid P-value* equals *half* the probability of the observed result, plus the probability of more extreme results. It has a null expected value of .5, the same as the regular P-value for continuous variates. For the tea-tasting data, with an observed value of 3 for n_{11}, the one-sided mid P-value equals $P(3)/2 + P(4) = .229/2 + .014 = .129$, compared to .243 for the ordinary P-value. The mid P-value for the two-sided test based on the X^2 statistic equals $P(X^2 = 2)/2 + P(X^2 = 8) = .257$, compared to .486 for the ordinary P-value.

Unlike an exact test with ordinary P-value, a test using the mid P-value does not guarantee that the Type I error rate falls below a fixed value (see Problem 2.27). However, it usually performs well and is less conservative than Fisher's exact test. For either P-value, rather than reducing the data to the extreme binary decision (reject

H_0, do not reject H_0), it is better simply to report the P-value, using it as a measure of the weight of evidence against the null hypothesis.

In Table 2.8, both margins are naturally fixed. When only one set is fixed, such as when rows totals are fixed with independent binomial samples, alternative exact tests exist that are less conservative than Fisher's exact test. These are beyond the scope of this text, but the reader can refer to a recent article by R. Berger and D. Boos (*J. Am. Statist. Assoc.*, 1994, p. 1012).

2.6.4 Small-Sample Confidence Interval for Odds Ratio

Exact inference is not limited to testing. One can also construct small-sample confidence intervals for the odds ratio. They correspond to a generalization of Fisher's exact test that tests an arbitrary value, $H_0 : \theta = \theta_0$. A 95% confidence interval contains all values of θ_0 for which the exact test of $H_0 : \theta = \theta_0$ yields $P > .05$; that is, for which one would not reject the null hypothesis at the .05 level.

As happens with exact tests, the discreteness makes these confidence intervals conservative. The true confidence level can be no smaller than the nominal one, but it may actually be considerably larger. For instance, a nominal 95% confidence interval may have true confidence level 98%. Moreover, the true level is unknown. The difference between the nominal and true levels can be considerable when the sample size is small. To reduce the conservativeness, one can construct the interval corresponding to the test using a mid P-value. The confidence interval consists of all θ_0 values for which the mid P-value exceeds .05. This interval is shorter. Though its actual confidence level is not guaranteed to be at least the nominal level, it tends to be close to that level. Computations for either of these types of confidence intervals are complex and require specialized software (e.g., StatXact, Cytel Software, Cambridge, MA).

For the tea-tasting data (Table 2.8), the "exact" 95% confidence interval for the true odds ratio equals (0.21, 626.17). The interval based on the test using the mid P-value equals (0.31, 308.55). Both intervals are very wide, because the sample size is so small.

2.6.5 Exact Tests of Independence for Larger Tables*

Exact tests of independence for tables of size larger than 2×2 use a multivariate version of the hypergeometric distribution. This distribution also applies to the set of all tables having the same row and column margins as the observed table. The exact tests are not practical to compute by hand or calculator but are feasible using computers. One selects a test statistic that describes the distance of the observed data from H_0. One then computes the probability of the set of tables for which the test statistic is at least as great as the observed one. For instance, for nominal variables, one could use X^2 as the test statistic. The P-value is then the null probability that X^2 is at least as large as the observed value, the calculation being done using the exact distribution rather than the large-sample chi-squared distribution.

Table 2.10 Example of 3 × 9 Table for Small-Sample Test

0	7	0	0	0	0	0	1	1
1	1	1	1	1	1	1	0	0
0	8	0	0	0	0	0	0	0

Recently developed software makes exact tests feasible for tables for which large-sample approximations are invalid. The software StatXact performs many exact inferences for categorical data. To illustrate, Table 2.10 is a 3×9 table having many zero entries and small counts. For it, $X^2 = 22.3$ with $df = 16$. The chi-squared approximation for the distribution of X^2 gives $P = .13$. Because the cell counts are so small, the validity of this approximation is suspect. Using StatXact to generate the exact sampling distribution of X^2, we obtain an exact P-value of .001, quite different from the result using the large-sample approximation.

For another example, we return to the analysis in Section 2.5 of Table 2.7, on the potential effect of maternal alcohol consumption on infant sex organ malformation. For testing independence, the values of $X^2 = 12.1$ and $G^2 = 6.2$ yield P-values from a chi-squared distribution with $df = 4$ of .02 and .19, respectively. Because of the imbalance in the table counts and the presence of some small counts, we could instead use exact tests for these statistics. The P-values using the exact distributions of X^2 and G^2 are .03 and .13, respectively. These are closer together but still give differing evidence about the association.

The columns of Table 2.7 are ordinal, and Section 2.5 presented a large-sample ordinal test based on a statistic M^2 (formula 2.5.1) that assigns scores to rows and columns. For ordinal data, exact tests exist using this statistic or using M for one-sided alternatives. For the one-sided alternative of a positive association, the exact P-value equals .02 for the midpoint scores (0, 0.5, 1.5, 4, 7), .10 for the equally spaced scores (0, 1, 2, 3, 4) and .29 for the midrank scores. For these data, the result depends greatly on the choice of scores.

PROBLEMS

2.1. A Swedish study considered the effect of low-dose aspirin on reducing the risk of stroke and heart attacks among people who have already suffered a stroke (*Lancet 338*: 1345–1349 (1991)). Of 1360 patients, 676 were randomly assigned to the aspirin treatment (one low-dose tablet a day) and 684 to a placebo treatment. During a follow-up period averaging about three years, the number of deaths due to myocardial infarction were 18 for the aspirin group and 28 for the placebo group.

a. Calculate and interpret the difference of proportions, relative risk of death, and the odds ratio.

b. Conduct an inferential analysis for these data. Interpret results.

2.2. In the United States, the estimated annual probability that a woman over the age of 35 dies of lung cancer equals .001304 for current smokers and .000121 for