CHAPTER 8

# Analyzing Association between Categorical Variables



Recall that we say there is an ***association*** between two variables if the distribution of the response variable changes in some way as the value of the explanatory variable changes. In comparing two groups, an association exists if the population means or population proportions differ between the groups.

This chapter presents methods for detecting and describing associations between two categorical variables. The methods of this chapter help us answer a question such as, "Is there an association between happiness and whether one is religious?" The methods of Chapter 7 for comparing two proportions are special cases of ones considered here in which both variables have only two categories.

Section 8.1 introduces terminology for categorical data analysis and defines *statistical independence*, a type of lack of association. Section 8.2 presents a significance test for determining whether two categorical variables are associated, and Section 8.3 follows up that test by a *residual analysis* that describes the nature of that association. Section 8.4 shows how to determine whether the association is strong enough to have practical importance. Sections 8.5 and 8.6 present specialized analyses for ordinal variables.

## 8.1 CONTINGENCY TABLES

Data for the analysis of categorical variables are displayed in ***contingency tables***. This type of table displays the number of subjects observed at all combinations of possible outcomes for the two variables.

### EXAMPLE 8.1 Gender Gap in Political Beliefs

In recent years in the United States political commentators have discussed whether a "gender gap" exists in political beliefs. Do women and men tend to differ in their political thinking and voting behavior? To investigate this, we study Table 8.1, from the 2004 GSS. The categorical variables are gender and political party identification (SEX and PARTYID in the GSS). Subjects indicated whether they identified more strongly with the Democratic or Republican party or as Independents.

Table 8.1 contains responses for 2771 subjects, cross-classified by their gender and party ID. Table 8.1 is called a $2 \times 3$ (read "2-by-3") contingency table, meaning that

From Chapter 8 of *Statistical Methods for the Social Sciences*, Fourth Edition. Alan Agresti, Barbara Finlay.

**TABLE 8.1:** Party Identification (ID) and Gender, for GSS Data

| | Party Identification | | | |
|---|---|---|---|---|
| Gender | Democrat | Independent | Republican | Total |
| Females | 573 | 516 | 422 | 1511 |
| Males | 386 | 475 | 399 | 1260 |
| Total | 959 | 991 | 821 | 2771 |

it has two rows and three columns. The row totals and the column totals are called the ***marginal distributions***. The sample marginal distribution for party identification, for instance, is the set of marginal frequencies (959, 991, 821). ■

## Percentage Comparisons

Constructing a contingency table from a data file is the first step in investigating an association between two categorical variables. To study how party identification depends on gender, we convert the frequencies to percentages within each row, as Table 8.2 shows. For example, a proportion of 573/1511 = 0.38, or 38% in percentage terms, identify themselves as Democrat. The percentage of males who identify themselves as Democrat equals 31% (386 out of 1260). It seems that females are more likely than males to identify as Democrats.

**TABLE 8.2:** Party Identification and Gender: Percentages Computed within Rows of Table 8.1

| | Party Identification | | | | |
|---|---|---|---|---|---|
| Gender | Democrat | Independent | Republican | Total | $n$ |
| Females | 38% | 34% | 28% | 100% | 1511 |
| Males | 31% | 38% | 32% | 101% | 1260 |

The two sets of percentages for females and males are called the ***conditional distributions*** on party identification. They refer to the sample data distribution of party ID, *conditional* on gender. The females' conditional distribution on party ID is the set of percentages (38, 34, 28) for (Democrat, Independent, Republican). The percentages sum to 100 in each row, except possibly for rounding. Figure 8.1 portrays graphically the two conditional distributions.

In a similar way, we could compute conditional distributions on gender for each party ID. The first column would indicate that 60% of the Democrats are females and 40% are males. In practice, it is standard to form the conditional distribution for the response variable, within categories of the explanatory variable. In this example, party ID is a response variable, so Table 8.2 reports percentages within rows, which tells us the percentage of (Democrats, Independents, Republicans) for each gender.

Another way to report percentages provides a single set for all cells in the table, using the total sample size as the base. To illustrate, in Table 8.1, of the 2771 subjects, 573 or 21% fall in the cell (Female, Democrat), 386 or 14% fall in the cell (Male, Democrat), and so forth. This percentage distribution is called the sample ***joint distribution***. It is useful for comparing relative frequencies of occurrences for combinations of variable levels. When we distinguish between response and explanatory variables, though, conditional distributions are more informative than the joint distribution.

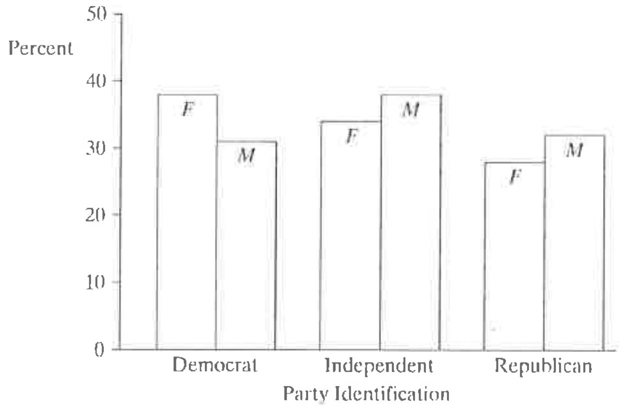


**FIGURE 8.1:** Portrayal of Conditional Distributions on Party ID in Table 8.2 for Females and Males

## Guidelines in Forming Contingency Tables

Here are some guidelines when finding proportions or percentages in contingency tables. First, as just noted, find them for the response variable within the categories of the explanatory variable. We'll construct the table so that the column variable is the response variable, as in Table 8.1. So we find proportions within each row, dividing each cell count by the row total.

Second, clearly label the variables and their categories, and give the table a title that identifies the variables and other relevant information. Third, include the total sample sizes on which the percentages or proportions are based. That way, readers can determine the cell frequencies, if they are not listed, and they can find standard errors to analyze the precision of sample proportion estimates.

## Independence and Dependence

Whether an association exists in Table 8.1 is a matter of whether females and males differ in their conditional distributions on party ID. We answer the question "Is party ID associated with gender?" with reference to the concepts of statistical *independence* and *dependence*.

**Statistical Independence and Statistical Dependence**

Two categorical variables are ***statistically independent*** if the population conditional distributions on one of them are identical at each category of the other. The variables are ***statistically dependent*** if the conditional distributions are not identical.

In other words, two variables are statistically independent if the percentage of the population in any particular category of one variable is the same for all categories of the other variable. In Table 8.2, the two conditional distributions are not identical. But that table describes a sample, and the definition of statistical independence refers to the population. If those observations were the entire population, then the variables would be statistically dependent.

For simplicity, we usually use the term *independent* rather than *statistically independent*. Table 8.3 is a hypothetical contingency table showing independence. The table contains the population data for two variables—party ID and ethnic group.

**TABLE 8.3:** Population Cross-Classification Exhibiting Statistical Independence. The conditional distribution is the same in each row, (44%, 14%, 42%).

| | Party Identification | | | |
|---|---|---|---|---|
| Ethnic Group | Democrat | Independent | Republican | Total |
| White | 440 (44%) | 140 (14%) | 420 (42%) | 1000 (100%) |
| Black | 44 (44%) | 14 (14%) | 42 (42%) | 100 (100%) |
| Hispanic | 110 (44%) | 35 (14%) | 105 (42%) | 250 (100%) |

The percentage of Democrats is the same for each ethnic group, 44%. Similarly, the percentage of Independents and the percentage of Republicans is the same for each ethnic group. The probability that a person has a particular party ID is the same for each ethnic group, and so party ID is independent of ethnic group.

Statistical independence is a symmetric property between two variables: If the conditional distributions within rows are identical, then so are the conditional distributions within columns. In Table 8.3, for example, you can check that the conditional distribution within each column equals (74%, 7%, 19%).

### EXAMPLE 8.2 What's Associated with Belief in Life after Death?

In recent General Social Surveys, the percentage of Americans who express a belief in life after death (variable AFTERLIF in the GSS) has been about 80%. This has been true both for females and for males and true for those who classify their race as black, white, or other. Thus, it appears that belief in life after death may be statistically independent of variables such as gender and race. On the other hand, whereas about 80% of Catholics and Protestants believe in an afterlife, only about 40% of Jews and 50% of those with no religion believe in an afterlife. We can't be sure, not having data for the entire population, but it seems that belief in life after death and religion are statistically dependent. ■

## 8.2 CHI-SQUARED TEST OF INDEPENDENCE

Table 8.1 contains sample data. The definition of statistical independence refers to the population. Two variables are independent if the *population* conditional distributions on the response variable are identical. Since Table 8.1 refers to a sample, it provides evidence but does not definitively answer whether party ID and gender are independent. Even if they are independent, we would not expect the *sample* conditional distributions to be identical. Because of sampling variability, we expect sample percentages to differ from the population percentages.

We next study whether it is plausible that party ID and gender are independent. If they are truly independent, could we expect sample differences such as Table 8.2 shows between females and males in their conditional distributions merely by sampling variation? Or would differences of this size be unlikely? To address this with a significance test, we test the following:

$H_0$: The variables are statistically independent.
$H_a$: The variables are statistically dependent.

The test requires randomization—for example, random sampling or a randomized experiment. The sample size must be large, satisfying a condition stated later in the section.

## Expected Frequencies for Independence

The chi-squared test compares the observed frequencies in the contingency table with values that satisfy the null hypothesis of independence. Table 8.4 shows the observed frequencies from Table 8.1, with the values (in parentheses) that satisfy $H_0$. These $H_0$ values have the same row and column totals as the observed frequencies, but satisfy independence. They are called ***expected frequencies***.

**TABLE 8.4:** Party Identification by Gender, with Expected Frequencies in Parentheses

| | Party Identification | | | |
|---|---|---|---|---|
| Gender | Democrat | Independent | Republican | Total |
| Female | 573 (522.9) | 516 (540.4) | 422 (447.7) | 1511 |
| Male | 386 (436.1) | 475 (450.6) | 399 (373.3) | 1260 |
| Total | 959 | 991 | 821 | 2771 |

**Observed and Expected Frequencies**

Let $f_o$ denote an ***observed*** frequency in a cell of the table. Let $f_e$ denote an ***expected*** frequency. This is the count expected in a cell if the variables were independent. It equals the product of the row and column totals for that cell, divided by the total sample size.

For instance, the cell in the upper left-hand corner refers to Females who identify as Democrats. For it, $f_o = 573$. Its expected frequency is $f_e = (1511)(959)/2771 = 522.9$, the product of the row total for Females and the column total for Democrats, divided by the overall sample size.

Let's see why this rule makes sense. In the entire sample, 959 out of 2771 people (34.6%) identify as Democrats. If the variables were independent, we would expect 34.6% of males and 34.6% of females to identify as Democrats. For instance, 34.6% of the 1511 Females should be classified in the Democrat category. The expected frequency for the cell is then

$$f_e = \left(\frac{959}{2771}\right)1511 = 0.346(1511) = 522.9.$$

## Chi-Squared Test Statistic

The test statistic for $H_0$: independence summarizes how close the expected frequencies fall to the observed frequencies. Symbolized by $\chi^2$, it is called the ***chi-squared statistic***. It equals

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}.$$

The summation is taken over all cells in the contingency table. For each cell, we square the difference between the observed and expected frequencies and then divide that square by the expected frequency. This is the oldest test statistic in use today; it was introduced by the British statistician Karl Pearson in 1900.

When $H_0$ is true, $f_o$ and $f_e$ tend to be close for each cell, and $\chi^2$ is relatively small. If $H_0$ is false, at least some $f_o$ and $f_e$ values tend not to be close, leading to large $(f_o - f_e)^2$ values and a large test statistic. The larger the $\chi^2$ value, the greater the evidence against $H_0$: independence.

Substituting the $f_o$ and $f_e$ values from Table 8.2 into the formula for $\chi^2$, we get

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$
$$= \frac{(573 - 522.9)^2}{522.9} + \frac{(516 - 540.4)^2}{540.4} + \frac{(422 - 447.7)^2}{447.7}$$
$$+ \frac{(386 - 436.1)^2}{436.1} + \frac{(475 - 450.6)^2}{450.6} + \frac{(399 - 373.3)^2}{373.3}$$
$$= 4.8 + \cdots + 1.8 = 16.2.$$

The calculation is messy, but it is simple to get $\chi^2$ using software. We next study how to interpret its magnitude.

### The Chi-Squared Distribution

The sampling distribution of the $\chi^2$ test statistic indicates how large $\chi^2$ must be before strong evidence exists that $H_0$ is false. For large sample sizes, the sampling distribution is the ***chi-squared probability distribution***. The name of the test and the symbol for the test statistic refer to the name of the sampling distribution. Here are the main properties of the chi-squared distribution:

- It is concentrated on the positive part of the real line. The $\chi^2$ test statistic cannot be negative, since it sums squared differences divided by positive expected frequencies. The minimum possible value, $\chi^2 = 0$, would occur if $f_o = f_e$ in each cell.
- It is skewed to the right.
- The precise shape of the distribution depends on the ***degrees of freedom*** ($df$). The mean $\mu = df$ and the standard deviation $\sigma = \sqrt{2df}$. Thus, the distribution tends to shift to the right and become more spread out for larger $df$ values. In addition, as $df$ increases, the skew lessens and the chi-squared curve becomes more bell shaped. See Figure 8.2.

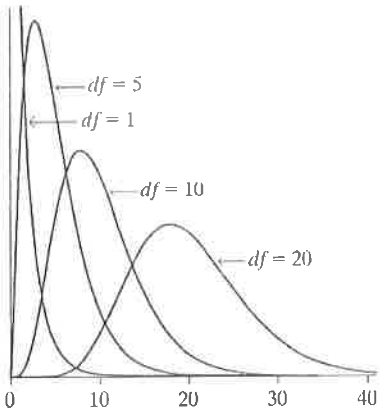


FIGURE 8.2: The Chi-Squared Distribution. The curve has larger mean and standard deviation as the degrees of freedom increase.

- For testing $H_0$: independence with a table having $r$ rows and $c$ columns,

$$df = (r - 1)(c - 1).$$

For a $2 \times 3$ table, $r = 2$ and $c = 3$ and $df = (2 - 1)(3 - 1) = 1 \times 2 = 2$. Larger numbers of rows and columns produce larger $df$ values. Since larger tables have more terms in the summation for the $\chi^2$ test statistic, the $\chi^2$ values also tend to be larger.

- The larger the $\chi^2$ value, the stronger the evidence against $H_0$: independence. The $P$-value equals the right-tail probability above the observed $\chi^2$ value. It measures the probability, presuming $H_0$ is true, that $\chi^2$ is at least as large as the observed value. Figure 8.3 depicts the $P$-value.

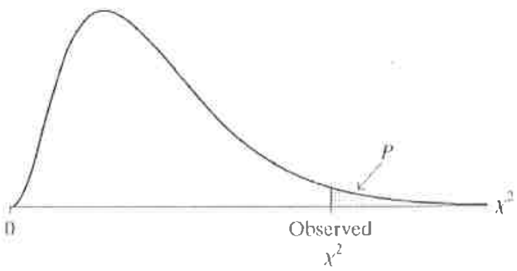


FIGURE 8.3: The $P$-Value for the Chi-Squared Test of Independence Is the Right-Tail Probability, above the Observed Value of the Test Statistic

Table C at the back of the text lists chi-squared values for various right-tail probabilities. These are $\chi^2$ test statistic values that have $P$-values equal to those probabilities. For example, Table C reports that when $df = 2$, $\chi^2 = 5.99$ has $P$-value $= 0.05$, and $\chi^2 = 9.21$ has $P$-value $= 0.01$.

### EXAMPLE 8.3 Chi-Squared Statistic for Party ID and Gender

To apply the chi-squared test to Table 8.4, we test the following:

$H_0$: Party ID and gender are statistically independent.
$H_a$: Party ID and gender are statistically dependent.

Previously, we obtained test statistic $\chi^2 = 16.2$. In Table C, for $df = 2$, 16.2 falls above 13.82, the chi-squared value having right-tail probability 0.001. Thus, we conclude that $P < 0.001$. Software indicates that $P = 0.0003$. This provides extremely strong evidence against $H_0$. It seems likely that party ID and gender are associated in the population. If the variables were independent, it would be highly unusual for a random sample to have this large a $\chi^2$ statistic. ■

### Sample Size Requirements

The chi-squared test, like one- and two-sample $z$ tests for proportions, is a large-sample test. The chi-squared distribution is the sampling distribution of the $\chi^2$ test statistic only if the sample size is large. A rough guideline for this requirement is that the expected frequency $f_e$ should exceed 5 in each cell. Otherwise, the chi-squared distribution may poorly approximate the actual distribution of the $\chi^2$ statistic.

For 2 × 2 contingency tables, a small-sample test of independence is ***Fisher's exact test***, discussed in Section 7.5. This test extends to tables of arbitrary size $r \times c$ but requires specialized software, such as SAS (the EXACT option in PROC FREQ) or SPSS (the Exact module).

If you have such software, you can use the exact test for *any* sample size. You don't need to use the chi-squared approximation. For Table 8.4, the exact test also gives $P$-value = 0.0003. Table 8.5 summarizes the five parts of the chi-squared test.

**TABLE 8.5:** The Five Parts of the Chi-Squared Test of Independence

1. Assumptions: Two categorical variables, random sampling, $f_e \geq 5$ in all cells
2. Hypotheses: $H_0$: Statistical independence of variables
   $H_a$: Statistical dependence of variables
3. Test statistic: $\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$, where $f_e = \frac{\text{(Row total)(Column total)}}{\text{Total sample size}}$
4. $P$-value: $P$ = right-tail probability above observed $\chi^2$ value,
   for chi-squared distribution with $df = (r - 1)(c - 1)$
5. Conclusion: Report $P$-value
   If decision needed, reject $H_0$ at $\alpha$-level if $P \leq \alpha$

### Using Software to Conduct Chi-Squared Tests

The chi-squared test of independence is computationally messy enough that you should use software to conduct it. Table 8.6 illustrates output for Table 8.1. SPSS lists the $P$-value under *Asymp. Sig.*, short for *asymptotic significance*, where the "asymptotic" refers to a large-sample method. Most software also reports an alternative test statistic, called the *likelihood-ratio statistic*, which usually provides similar results. Chapter 15 introduces this statistic.

**TABLE 8.6:** Printout for Chi-Squared Test of Independence

```
GENDER                              PARTYID
                          |democrat|  indep |  repub |  Total
--------------------------+--------+--------+--------+
female  Count             |    573 |    516 |    422 |   1511
        Expected Count    |  522.9 |  540.4 |  447.7 |
--------------------------+--------+--------+--------+
male    Count             |    386 |    475 |    399 |   1260
       Expected Count     |  436.1 |  450.6 |  373.3 |
--------------------------+--------+--------+--------+
Total                          959      991      821     2771

Statistic                Value      df       Asymp. Sig.
----------------------------------------------------------
Pearson Chi-Square      16.202       2           .000
Likelihood Ratio        16.273       2           .000
```

### Interpretation of Degrees of Freedom

The $df$ in a chi-squared test has the following interpretation: Given the marginal totals, the cell counts in a rectangular block of size $(r - 1) \times (c - 1)$ within the contingency table determine the other cell counts.

To illustrate, in Table 8.1, suppose we know the two frequencies 573 and 516 in the upper-left-hand part of the table. This is a block of size $1 \times 2$, shown in Table 8.7. Then, given the marginal totals, we can determine all the other cell counts. For instance, since 573 of the 959 Democrats are female, the other $959 - 573 = 386$ must be male. Since 516 of the 991 Independents are female, the other $991 - 516 = 475$ must be male. Also, since the total of the female row is 1511, and since the first two cells contain 1089 (i.e., $573 + 516$) subjects, the remaining cell must have $1511 - 1089 = 422$ observations. From this and the fact that the last column has 821 observations, there must be $821 - 422 = 399$ observations in the second cell in that column.

**TABLE 8.7:** Illustration of Degrees of Freedom; a Block of $(r - 1)(c - 1)$ Cell Counts Determine the Others

| | Party Identification | | | |
|---|---|---|---|---|
| Gender | Democrat | Independent | Republican | Total |
| Female | 573 | 516 | — | 1511 |
| Male | — | — | — | 1260 |
| Total | 959 | 991 | 821 | 2771 |

Once the marginal frequencies are fixed in a contingency table, a block of only $(r - 1) \times (c - 1)$ cell counts is free to vary, since these cell counts determine the remaining ones. The degrees of freedom value equals the number of cells in this block, so $df = (r - 1)(c - 1)$. We'll see another way to interpret $df$ at the end of Section 8.3.

### Chi-Squared Tests and Treatment of Categories

In the chi-squared test, the value of the $\chi^2$ test statistic does not depend on which is the response variable and which is the explanatory variable (if either). The steps of the test and the results are identical either way. When a response variable is identified and the population conditional distributions are identical, they are said to be *homogeneous*. The chi-squared test of independence is then often referred to as a *test of homogeneity*. For example, party ID is a response variable and gender is explanatory, so we can regard the chi-squared test applied to these data as a test of homogeneity of the conditional distributions of party ID.

The chi-squared test treats the classifications as nominal. That is, $\chi^2$ takes the same value if the rows or columns are reordered in any way. If either classification is ordinal or grouped interval, the chi-squared test does not use that information. In that case, it is usually better to apply stronger statistical methods designed for the higher level of measurement. Section 8.6 presents a test of independence for ordinal variables.

## 8.3 RESIDUALS: DETECTING THE PATTERN OF ASSOCIATION

The chi-squared test of independence, like other significance tests, provides limited information. If the $P$-value has moderate size (e.g., $P > 0.10$), it is plausible that the variables are independent. If the $P$-value is very small, strong evidence exists that the variables are associated. The chi-squared test tells us nothing, however, about the nature or strength of the association. The test does not indicate whether all cells deviate greatly from independence or perhaps only one or two of the cells do so. The next two sections introduce methods to learn more about the association.

## Residual Analysis

A cell-by-cell comparison of observed and expected frequencies reveals the nature of the evidence about the association. The difference $(f_o - f_e)$ between an observed and expected cell frequency is called a ***residual***.

The counts on party ID and gender are shown again below in Table 8.8. For the first cell, the residual equals $573 - 522.9 = 50.1$. The residual is positive when, as in this cell, the observed frequency $f_o$ exceeds the value $f_e$ that independence predicts. The residual is negative when the observed frequency is smaller than independence predicts.

How do we know whether a residual is large enough to indicate a departure from independence that is unlikely to be due to mere chance? A standardized form of the residual that behaves like a $z$-score provides this information.

**Standardized Residual**

The ***standardized residual*** for a cell equals

$$z = \frac{f_o - f_e}{se} = \frac{f_o - f_e}{\sqrt{f_e(1 - \text{row proportion})(1 - \text{column proportion})}}.$$

Here, $se$ denotes the standard error of $f_o - f_e$, presuming $H_0$ is true. The standardized residual is the number of standard errors that $(f_o - f_e)$ falls from the value of 0 that we expect when $H_0$ is true.

The $se$ uses the marginal proportions for the row and the column in which the cell falls. When $H_0$: independence is true, the standardized residuals have a large-sample standard normal distribution. They fluctuate around a mean of 0, with a standard deviation of about 1.

We use the standardized residuals in an informal manner to describe the pattern of the association among the cells. A large standardized residual provides evidence against independence in that cell. When $H_0$ is true, there is only about a 5% chance that any particular standardized residual exceeds 2 in absolute value. When we inspect many cells in a table, some standardized residuals could be large just by random variation. Values below $-3$ or above $+3$, however, are very convincing evidence of a true effect in that cell.

### EXAMPLE 8.4 Standardized Residuals for Gender and Political ID

Table 8.8 displays the standardized residuals for testing independence between gender and party affiliation. For the first cell, for instance, $f_o = 573$ and $f_e = 522.9$. The first row and first column marginal proportions equal $1511/2771 = 0.545$ and $959/2771 = 0.346$. Substituting into the formula, the standardized residual

$$z = \frac{f_o - f_e}{\sqrt{f_e(1 - \text{row prop.})(1 - \text{column prop.})}} = \frac{573 - 522.9}{\sqrt{[522.9(1 - 0.545)(1 - 0.346)]}} = 4.0.$$

Since the standardized residual exceeds 3.0, this cell has more observations than we'd expect if the variables were truly independent.

Table 8.8 exhibits very large positive residuals for female Democrats and male Republicans. This means there were more female Democrats and male Republicans than the hypothesis of independence predicts. The table exhibits relatively large negative residuals for female Republicans and male Democrats. There were fewer

**TABLE 8.8:** Standardized Residuals (in Parentheses) for Testing Independence between Party ID and Gender

| | Party Identification | | |
|---|---|---|---|
| Gender | Democrat | Independent | Republican |
| Female | 573 (4.0) | 516 (−1.9) | 422 (−2.1) |
| Male | 386 (−4.0) | 475 (1.9) | 399 (2.1) |

female Republicans and male Democrats than we'd expect if party affiliation were independent of gender.

For each party ID, Table 8.8 contains only one nonredundant standardized residual. The one for females is the negative of the one for males. The observed counts and the expected frequencies have the same row and column totals. Thus, in a given column, if $f_o > f_e$ in one cell, the reverse must happen in the other cell. The differences $f_o - f_e$ have the same magnitude but different sign in the two cells, implying the same pattern for their standardized residuals. ■

Along with the $\chi^2$ statistic, most statistical software can provide standardized residuals. See the text appendix for details.

### Chi-Squared and Difference of Proportions for 2 × 2 Tables

As Section 7.2 showed, 2 × 2 contingency tables often compare two groups on a binary response variable. The outcomes could be, for example, (yes, no) on an opinion question. For convenience, we label the two possible outcomes for that binary variable by the generic labels *success* and *failure*.

Let $\pi_1$ represent the proportion of successes in population 1, and let $\pi_2$ represent the proportion of successes in population 2. Then $(1 - \pi_1)$ and $(1 - \pi_2)$ are the proportions of failures. Table 8.9 displays the notation. The rows are the groups to be compared and the columns are the response categories.

**TABLE 8.9:** 2 × 2 Table for Comparing Two Groups on a Binary Response Variable

| | Proportion Making Each Response | | |
|---|---|---|---|
| Group | Success | Failure | Total |
| 1 | $\pi_1$ | $1 - \pi_1$ | 1.0 |
| 2 | $\pi_2$ | $1 - \pi_2$ | 1.0 |

If the response variable is statistically independent of the populations considered, then $\pi_1 = \pi_2$. The null hypothesis of independence corresponds to the *homogeneity* hypothesis, $H_0$: $\pi_1 = \pi_2$. In fact, the chi-squared test of independence is equivalent to a test for equality of two population proportions. Section 7.2 presented a $z$ test statistic for this, based on dividing the difference of sample proportions by its standard error,

$$z = \frac{\hat{\pi}_2 - \hat{\pi}_1}{se}.$$

The chi-squared statistic relates to this $z$ statistic by $\chi^2 = z^2$.

The chi-squared statistic for 2 × 2 tables has $df = 1$. Its $P$-value from the chi-squared distribution is the same as the $P$-value for the two-sided test with the $z$ test statistic. This is because of a direct connection between the standard normal distribution and the chi-squared distribution with $df = 1$: Squaring $z$-scores with certain two-tail probabilities yields chi-squared scores with $df = 1$ having the same right-tail probabilities. For instance, $z = 1.96$ is the $z$-score with a two-tail probability of 0.05. The square of this, $(1.96)^2 = 3.84$, is the chi-squared score for $df = 1$ with a $P$-value of 0.05. (You can check this in Table C.)

**EXAMPLE 8.5 Women's and Men's Roles**

Table 8.10 summarizes responses from General Social Surveys in 1977 and in 2006 to the statement (FEFAM), "It is much better for everyone involved if the man is the achiever outside the home and the woman takes care of the home and family." You can check that the sample proportions agreeing with the statement were $\hat{\pi}_1 = 0.658$ in 1977, $\hat{\pi}_2 = 0.358$ in 2006, the *se* for the test comparing them equals 0.0171, and the $z$ test statistic for $H_0$: $\pi_1 = \pi_2$ is $z = (0.658 - 0.358)/0.0171 = 17.54$. You can also check that the chi-squared statistic for this table is $\chi^2 = 307.6$. This equals the square of the $z$ test statistic. Both statistics show extremely strong evidence against the null hypothesis of equal population proportions. ■

**TABLE 8.10:** GSS Responses to the Statement, "It is much better for everyone involved if the man is the achiever outside the home and the woman takes care of the home and family," with Standardized Residuals in Parentheses

| Year | Agree | Disagree | Total |
|---|---|---|---|
| 1977 | 989 (17.5) | 514 (−17.5) | 1503 |
| 2006 | 704 (−17.5) | 1264 (17.5) | 1968 |

## Standardized Residuals for 2 × 2 Tables

Let's follow up the test for Table 8.10 with a residual analysis. Table 8.10 also shows the standardized residuals. Those in the first column suggest that more subjects agreed with the statement in 1977 and fewer agreed in 2006 than we'd expect if opinion were independent of the year of the survey. Notice that *every* standardized residual equals either +17.5 or −17.5. The absolute value of the standardized residual is 17.5 in every cell.

For chi-squared tests with 2 × 2 tables, $df = 1$. This means that only one piece of information exists about whether an association exists. Once we find the standardized residual for one cell, other standardized residuals in the table have the same absolute value. In fact, in 2 × 2 tables, each standardized residual equals the $z$ test statistic (or its negative) for comparing the two proportions. The square of each standardized residual equals the $\chi^2$ test statistic.

## Chi-Squared Needed for Larger Tables Than 2 × 2

For a 2 × 2 table, why should we ever do a $z$ test if we can get the same result with chi-squared? An advantage of the $z$ test is that it also applies with one-sided alternative hypotheses, such as $H_a$: $\pi_1 > \pi_2$. The direction of the effect is lost in squaring $z$ and using $\chi^2$.

Why do we need the $\chi^2$ statistic? The reason is that a $z$ statistic can only compare a single estimate to a single $H_0$ value. Examples are a $z$ statistic for comparing a sample proportion to a $H_0$ proportion such as 0.5, or a difference of sample proportions to a $H_0$ value of 0 for $\pi_2 - \pi_1$. When a table is larger than $2 \times 2$ and thus $df > 1$, we need more than one difference parameter to describe the association. For instance, suppose Table 8.10 had three rows, for three years of data. Then $H_0$: independence corresponds to $\pi_1 = \pi_2 = \pi_3$, where $\pi_i$ is the population proportion agreeing with the statement in year $i$. The comparison parameters are $(\pi_1 - \pi_2)$, $(\pi_1 - \pi_3)$, and $(\pi_2 - \pi_3)$. We could use a $z$ statistic for each comparison, but not a single $z$ statistic for the overall test of independence.

We can interpret the $df$ value in a chi-squared test as the number of parameters needed to determine all the comparisons for describing the contingency table. For instance, for a $3 \times 2$ table for comparing three years on a binary opinion response, $df = 2$. This means we need to know only two parameters for making comparisons to figure out the third. For instance, if we know $(\pi_1 - \pi_2)$ and $(\pi_1 - \pi_3)$, then

$$(\pi_2 - \pi_3) = (\pi_1 - \pi_3) - (\pi_1 - \pi_2).$$

## 8.4 MEASURING ASSOCIATION IN CONTINGENCY TABLES

The main questions normally addressed in analyzing a contingency table are as follows:

- *Is there an association?* The chi-squared test of independence addresses this question. The smaller the $P$-value, the stronger the evidence of association.
- *How do the data differ from what independence predicts?* The standardized residuals highlight the cells that are more likely or less likely than expected under independence.
- *How strong is the association?* To summarize this, we use a statistic such as a difference of proportions, forming a confidence interval to estimate the strength of association in the population.

Analyzing the *strength* of the association reveals whether the association is important or whether it is statistically significant but practically insignificant. This section presents two ways to measure strength of association for contingency tables.

### Measures of Association

**Measure of Association**

A ***measure of association*** is a statistic or a parameter that summarizes the strength of the dependence between two variables.

Let's first consider what is meant by *strong* versus *weak* association. Table 8.11 shows two hypothetical contingency tables relating race to opinion about allowing civil unions for same-sex couples. Case A, which exhibits statistical independence, represents the weakest possible association. Both whites and blacks have 60% in favor and 40% opposed to civil unions. Opinion is not associated with race. By contrast, case B exhibits the strongest possible association. All whites favor allowing civil unions, whereas all blacks oppose it. In this table, opinion is completely dependent on race. For these subjects, if we know their race, we know their opinion.

A measure of association describes how similar a table is to the tables representing the strongest and weakest associations. It takes a range of values from one extreme

TABLE 8.11: Cross-Classification of Opinion about Same-Sex Civil Unions, by Race, Showing (A) No Association, (B) Maximum Association

| Case A | Opinion | | | Case B | Opinion | | |
|---|---|---|---|---|---|---|---|
| Race | Favor | Oppose | Total | | Favor | Oppose | Total |
| White | 360 | 240 | 600 | | 600 | 0 | 600 |
| Black | 240 | 160 | 400 | | 0 | 400 | 400 |
| Total | 600 | 400 | 1000 | | 600 | 400 | 1000 |

to another as data range from the weakest to strongest association. It is useful for comparing associations, to determine which is stronger.

### Difference of Proportions

As Sections 7.2 and 8.3 discussed, many $2 \times 2$ tables compare two groups on a binary variable. In such cases, an easily interpretable measure of association is the difference between the proportions for a given response category. For example, we could measure the difference between the proportions of whites and blacks who favor allowing same-sex civil unions. For Table 8.11(A), this difference is

$$\frac{360}{600} - \frac{240}{400} = 0.60 - 0.60 = 0.0.$$

The population difference of proportions is 0 whenever the conditional distributions are identical, that is, when the variables are independent. The difference is 1 or $-1$ for the strongest possible association. For Table 8.11(B), for instance, the difference is

$$\frac{600}{600} - \frac{0}{400} = 1.0,$$

the maximum possible absolute value for the difference.

This measure falls between $-1$ and $+1$. In practice we don't expect data to take these extreme values, but *the stronger the association, the larger the absolute value of the difference of proportions.* The following contingency tables illustrate the increase in this measure as the degree of association increases:

| | | | | | | |
|---|---|---|---|---|---|---|
| Cell Counts: | 25 25 | 30 20 | 35 15 | 40 10 | 45 5 | 50 0 |
| | 25 25 | 20 30 | 15 35 | 10 40 | 5 45 | 0 50 |
| Difference of Proportions: | 0 | .2 | .4 | .6 | .8 | 1.0 |

For the second table, for instance, the proportion falling in the first column equals $30/(30 + 20) = 0.60$ in row 1 and $20/(20 + 30) = 0.40$ in row 2, for a difference of $0.60 - 0.40 = 0.20$.

### Chi-Squared Does Not Measure Association

A large value for $\chi^2$ in the test of independence suggests that the variables are associated. It does *not* imply, however, that the variables have a strong association. This statistic summarizes how close the observed frequencies are to the frequencies expected if the variables were independent. It merely indicates, however, how much

TABLE 8.12: Cross-Classifications of Opinion on Legalized Same-Sex Unions, by Race, Showing Weak but Identical Associations

| | A | | | B | | | C | | |
|---|---|---|---|---|---|---|---|---|---|
| | Yes | No | Total | Yes | No | Total | Yes | No | Total |
| White | 49 | 51 | 100 | 98 | 102 | 200 | 4,900 | 5,100 | 10,000 |
| Black | 51 | 49 | 100 | 102 | 98 | 200 | 5,100 | 4,900 | 10,000 |
| | 100 | 100 | 200 | 200 | 200 | 400 | 10,000 | 10,000 | 20,000 |
| | $\chi^2 = 0.08$ | | | $\chi^2 = 0.16$ | | | $\chi^2 = 8.0$ | | |
| | $P$-value = 0.78 | | | $P$-value = 0.69 | | | $P$-value = 0.005 | | |

evidence there is that the variables are dependent, not how strong that dependence is. For a given association, larger $\chi^2$ values occur for larger sample sizes. As with any significance test, large test statistic values can occur with weak effects, if the sample size is large.

For example, consider the hypothetical cases in Table 8.12. The association in each table is very weak—the conditional distribution for whites on opinion (49% favor, 51% oppose) is nearly identical to the conditional distribution for blacks (51% favor, 49% oppose). All three tables show exactly the same degree of association, with the difference between the proportions of blacks and whites who favor legalizing same-sex civil unions being $0.51 - 0.49 = 0.02$ in each table.

For the sample of size 200 in case A, $\chi^2 = 0.08$, which has a $P$-value $= 0.78$. For the sample of size 400 in case B, $\chi^2 = 0.16$, for which $P = 0.69$. So, when the cell counts double, $\chi^2$ doubles. Similarly, for the sample size of 20,000 (100 times as large as $n = 200$) in case C, $\chi^2 = 8.0$ (100 times as large as $\chi^2 = 0.08$), and $P = 0.005$.

In summary, for a fixed percentage assignment to the cells of a contingency table, $\chi^2$ is directly proportional to the sample size—larger values occur with larger sample sizes. Like other test statistics, the larger the $\chi^2$ statistic, the smaller the $P$-value and the stronger the evidence against the null hypothesis. However, a small $P$-value can result from a weak association when the sample size is large, as case C shows.

### The Odds Ratio*

The difference of proportions is easily interpretable. Several other measures are also reported by statistical software. This subsection presents the most important one for categorical data analysis, the *odds ratio*.

For a binary response variable, recall that we use *success* to denote the outcome of interest and *failure* the other outcome. The ***odds*** of success are defined to be

$$\text{Odds} = \frac{\text{Probability of success}}{\text{Probability of failure}}.$$

If the probability of success $= 0.75$, then the probability of failure equals $1 - 0.75 = 0.25$, and the odds of success $= 0.75/0.25 = 3.0$. If $P(\text{success}) = 0.50$, then odds $= 0.50/0.50 = 1.0$. If $P(\text{success}) = 0.25$, then odds $= 0.25/0.75 = 1/3$. The odds are nonnegative, with value greater than 1.0 when a success is more likely than a failure. When odds $= 3.0$, a success is three times as likely as a failure; we expect about three successes for every failure. When odds $= 1/3$, a failure is three times as likely as a success; we expect about one success for every three failures.

The probability of an outcome relates to the odds of the outcome by

$$\text{Probability} = \frac{\text{Odds}}{\text{Odds} + 1}.$$

For instance, when odds = 3, probability = $3/(3 + 1) = 0.75$.

The ratio of odds from the two rows of a $2 \times 2$ table is called the ***odds ratio***. For instance, if the odds = 4.5 in row 1 and the odds = 3.0 in row 2, then the odds ratio equals $4.5/3.0 = 1.5$. The odds of success in row 1 then equal 1.5 times the odds of success in row 2. We denote the odds ratio by the Greek letter $\theta$ (theta).

### EXAMPLE 8.6 Race of Murder Victims and Offenders

For murders in the United States in 2005 having a single victim and single offender, Table 8.13 cross classifies the race of the victim by the race of the offender. We treat race of victim as the response variable. For white offenders, the proportion of victims who were white equals $3150/3380 = 0.932$ and the proportion who were black equals $230/3380 = 0.068$. The odds of a white victim equaled $0.932/0.068 = 13.7$. This equals $(3150/3380)/(230/3380) = 3150/230$. So we can calculate the odds by the ratio of the counts in the two cells in row 1, without converting them to proportions.

**TABLE 8.13:** Cross-Classification of Race of Victim and Race of Offender

| Race of Offender | Race of Victim: White | Black | Total |
|---|---|---|---|
| White | 3150 | 230 | 3380 |
| Black | 516 | 2984 | 3500 |

*Source*: www.fbi.gov

The value 13.7 means that for white offenders, there were 13.7 white victims for every 1 black victim. For black offenders, the odds of a white victim equaled $516/2984 = 0.173$. This means there were 0.173 white victims for every 1 black victim. Equivalently, since $2984/516 = 1/0.173 = 5.8$, black offenders had 5.8 black victims for every white victim.

For Table 8.13, the odds ratio equals

$$\theta = \frac{\text{Odds for white offenders}}{\text{Odds for black offenders}} = \frac{13.7}{0.173} = 79.2.$$

For white offenders, the odds of a white victim were about 79 times the odds of a white victim for black offenders. ■

In summary,

**Odds and Odds Ratio**

The estimated ***odds*** for a binary response equal the number of successes divided by the number of failures.
The ***odds ratio*** is a measure of association for $2 \times 2$ contingency tables that equals the odds in row 1 divided by the odds in row 2.

#### Properties of the Odds Ratio*

In Table 8.13, suppose we treat race of offender, rather than race of victim, as the response variable. When victims were white, the odds the race of offender was white equaled $3150/516 = 6.10$. When victims were black, the odds the race of offender was white equaled $230/2984 = 0.077$. The odds ratio equals $6.10/0.077 = 79.2$. For each choice of the response variable, the odds ratio is 79.2. In fact,

- The odds ratio takes the same value regardless of the choice of response variable.

Since the odds ratio treats the variables symmetrically, the odds ratio is a natural measure when there is no obvious distinction between the variables, such as when both are response variables.

- The odds ratio $\theta$ equals the ratio of the products of cell counts from diagonally opposite cells.

For Table 8.13, for instance,

$$\theta = \frac{(3150 \times 2984)}{(230 \times 516)} = 79.2.$$

Because of this property, the odds ratio is also called the ***cross-product ratio***.

- The odds ratio can equal any nonnegative number.
- When the success probabilities are identical in the two rows of a $2 \times 2$ table (i.e., $\pi_1 = \pi_2$), then $\theta = 1$.

When $\pi_1 = \pi_2$, the odds are also equal. The odds of success do not depend on the row level of the table, and the variables are then independent, with $\theta = 1$. The value $\theta = 1$ for independence serves as a baseline for comparison. Odds ratios on each side of 1 reflect certain types of associations.

- When $\theta > 1$, 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.

- When $\theta < 1$, the odds of success are *lower* in row 1 than in row 2.
- Values of $\theta$ farther from 1.0 in a given direction represent stronger associations.

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 $\theta$ represent the same strength of association, but in opposite directions, when one value is the reciprocal of the other.

For instance, $\theta = 4.0$ and $\theta = 1/4.0 = 0.25$ represent the same strength of association. When $\theta = 0.25$, the odds of success in row 1 are 0.25 times the odds of success in row 2. Equivalently, the odds of success in row 2 are $1/0.25 = 4.0$ times the odds of success in row 1. When the order of the rows is reversed or the order of the columns is reversed, the new value of $\theta$ is the reciprocal of the original value. This ordering of rows or columns 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.

In interpreting the odds ratio, be careful not to misinterpret it as a ratio of probabilities. An odds ratio of 79.2 does *not* mean that $\pi_1$ is 79.2 times $\pi_2$. Instead, $\theta = 79.2$ means that the *odds* in row 1 equal 79.2 times the odds in row 2. The odds ratio is a ratio of two odds, not a ratio of two probabilities. That is,

$$\theta = \frac{\text{Odds in row 1}}{\text{Odds in row 2}} = \frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)}, \quad \text{not} \quad \frac{\pi_1}{\pi_2}.$$

The ratio $\pi_1/\pi_2$ is itself a useful measure. Section 7.1 introduced this measure, often called the ***relative risk***.

The sampling distribution of the sample odds ratio $\hat{\theta}$ is highly skewed unless the sample size is extremely large, in which case the distribution is approximately normal. See Exercise 8.45 for the method of constructing confidence intervals for odds ratios.

### Odds Ratios for $r \times c$ Tables*

For contingency tables with more than two rows or more than two columns, the odds ratio describes patterns in any $2 \times 2$ subtable. We illustrate using GSS data on political party ID and race, shown in Table 8.14.

**TABLE 8.14:** GSS Data from 2004 on Party Identification and Race

| | Party Identification | | |
|---|---|---|---|
| Gender | Democrat | Independent | Republican |
| Black | 250 | 106 | 17 |
| White | 640 | 783 | 775 |

Consider first the $2 \times 2$ subtable formed from the first two columns. The sample odds ratio equals $(250 \times 783)/(106 \times 640) = 2.89$. The odds that a black's response was Democrat rather than Independent equal 2.89 times the odds for whites. Of those subjects who responded Democrat or Independent, blacks were more likely than whites to respond Democrat.

The sample odds ratio for the last two columns of this table equals $(106 \times 775)/(17 \times 783) = 6.17$. The odds that a black's response was Independent rather than Republican equal 6.2 times the odds for whites. Of those subjects who responded Independent or Republican, blacks were much more likely than whites to respond Independent.

Finally, for the $2 \times 2$ subtable formed from the first and last columns, the sample odds ratio equals $(250 \times 775)/(17 \times 640) = 17.81$. The odds that a black's response was Democrat rather than Republican equal 17.8 times the odds for whites. Of those subjects who responded Democrat or Republican, blacks were much more likely than whites to respond Democrat. This is a very strong effect, far from the independence odds ratio value of 1.0.

The odds ratio value of 17.8 for the first and last columns equals (2.89)(6.17), the product of the other two odds ratios. For $2 \times 3$ tables, $df = 2$, meaning that only two bits of information exist about the association. Two of the odds ratios determine the third.

### Summary Measures of Association for $r \times c$ Tables*

Instead of studying association in $2 \times 2$ subtables, it's possible to summarize association in the entire table by a single number. One way to do this summarizes how well we can predict the value on one variable based on knowing the value on the other variable. For example, party ID and race are highly associated if race is a good predictor of party ID; that is, if knowing their race, we can make much better predictions about people's party ID than if we did not know it.

For quantitative variables, the *correlation* is such a summary measure. We'll study a similar summary measure of this type for ordinal variables (called *gamma*) in the next section. These measures describe an overall trend in the data. For nominal variables, when $r$ or $c$ exceed 2, it is usually an oversimplification to describe the table with a single measure of association. In that case, too many possible patterns of association exist to describe an $r \times c$ table well by a single number. Nominal measures based on predictive power (called *tau* and *lambda*) and gamma for ordinal data were defined in 1954 by two prominent statistician–social scientists, Leo Goodman and William Kruskal. Most software for analyzing contingency tables prints their measures and several others. Some nominal measures, such as the contingency coefficient and Cramer's $V$, are difficult to interpret (other than larger values representing stronger association) and, in our view, not especially useful.

We do not present the summary nominal measures in this text. We believe you get a better feel for the association by making percentage comparisons of conditional distributions, by viewing the pattern of standardized residuals in the cells of the table, by constructing odds ratios in $2 \times 2$ subtables, and by building models such as those presented in Chapter 15. These methods become even more highly preferred to summary measures of association when the analysis is multivariate rather than bivariate.

## 8.5 ASSOCIATION BETWEEN ORDINAL VARIABLES*

We now turn our attention to other analyses of contingency tables that apply when the variables are ordinal. The categories of ordinal variables are ordered. Statistical analyses for ordinal data take this ordering into account. This section introduces a popular ordinal measure of association, and Section 8.6 presents related methods of inference.

### EXAMPLE 8.7 How Strongly Associated Are Income and Happiness?

Table 8.15 is a contingency table with ordinal variables. These data, from the 2004 GSS, refer to the relation between family income (FINRELA) and happiness (HAPPY). This table shows results for black Americans, and Exercise 8.13 analyzes data for white Americans.

Let's first get a feel for the data by studying the conditional distributions on happiness. Table 8.15 shows these in parentheses. For instance, the conditional distribution (24%, 54%, 22%) displays the percentages in the happiness categories for subjects with family income below average. Only 22% are very happy, whereas 36% of the subjects at the highest income level are very happy. Conversely, a lower percentage (9%) of the high-income group are not too happy compared to those in the lowest income group (24%). The odds ratio for the four corner cells is $(16 \times 8)/(15 \times 2) = 4.3$. It seems that subjects with higher incomes tended to have greater happiness. ■

TABLE 8.15: Family Income and Happiness for a GSS Sample

| | Happiness | | | |
|---|---|---|---|---|
| Family Income | Not Too Happy | Pretty Happy | Very Happy | Total |
| Below average | 16 (24%) | 36 (54%) | 15 (22%) | 67 (100.0%) |
| Average | 11 (16%) | 36 (53%) | 21 (31%) | 68 (100.0%) |
| Above average | 2 (9%) | 12 (55%) | 8 (36%) | 22 (100.0%) |
| Total | 29 | 84 | 44 | 157 |

Ordinal data exhibit two primary types of association between variables $x$ and $y$—*positive* and *negative*. Positive association results when subjects at the high end of the scale on $x$ tend also to be high on $y$, and those who are low on $x$ tend to be low on $y$. For example, a positive association exists between income and happiness if those with low incomes tend to have lower happiness, and those with high incomes tend to have greater happiness. Negative association occurs when subjects classified high on $x$ tend to be classified low on $y$, and those classified low on $x$ tend to be high on $y$. For example, a negative association might exist between religious fundamentalism and tolerance toward homosexuality—the more fundamentalist in religious beliefs, the less tolerance toward homosexuality.

### Concordance and Discordance

Many ordinal measures of association are based on the information about the association provided by all the pairs of observations.

**Concordant Pair, Discordant Pair**

A pair of observations is ***concordant*** if the subject who is *higher* on one variable also is *higher* on the other variable.
A pair of observations is ***discordant*** if the subject who is *higher* on one variable is *lower* on the other.

In Table 8.15, we regard *Not too happy* as the low end and *Very happy* as the high end of the scale on $y$ = happiness, and *Below average* as low and *Above average* as high on $x$ = family income. By convention, we construct contingency tables for ordinal variables so that the low end of the row variable is the first row and the low end of the column variable is the first column. (There is no standard, however, and other books or software may use a different convention.)

Consider a pair of subjects, one of whom is classified (below average, not too happy), and the other of whom is classified (average, pretty happy). The first subject is one of the 16 classified in the upper-left-hand cell of Table 8.15, and the second subject is one of the 36 classified in the middle cell. This pair of subjects is concordant, since the second subject is higher than the first subject both in happiness and in income. The subject who is higher on one variable is also higher on the other. Now, each of the 16 subjects classified (below average, not too happy) can pair with each of the 36 subjects classified (average, pretty happy). So there are $16 \times 36 = 576$ concordant pairs of subjects from these two cells.

By contrast, each of 36 subjects in the cell (below average, pretty happy) forms a discordant pair when matched with each of the 11 subjects in the cell (average, not too happy). The 36 subjects have lower income than the other 11 subjects, yet they have greater happiness. All $36 \times 11 = 396$ of these pairs of subjects are discordant.

Concordant pairs of observations provide evidence of positive association since, for such a pair, the subject who is higher on one variable also is higher on the other. On the other hand, the more prevalent the discordant pairs, the more evidence there is of a negative association.

**Notation for Numbers of Concordant and Discordant Pairs**

Let $C$ denote the total number of concordant pairs of observations, and let $D$ denote the total number of discordant pairs of observations.

A general rule for finding the number of concordant pairs $C$ is this: Start at the corner of the table for the low level for each variable (the cell in row 1 and column 1 for Table 8.15). Multiply that cell count by the count in every cell that is higher on both variables (those cells below and to the right in Table 8.15). Similarly, for every other cell, multiply the cell count by the counts in cells that are higher on both variables. (For the cells in the row or in the column at the highest level of a variable, such as row *Above average* or column *Very happy* in Table 8.15, no observations are higher on both variables.) The number of concordant pairs is the sum of these products.

In Table 8.15, the 16 subjects in the first cell are concordant when matched with the (36 + 21 + 12 + 8) subjects below and to the right who are higher on each variable. Similarly, the 36 subjects in the second cell in the first row are concordant when matched with the (21 + 8) subjects who are higher on each variable, and so forth. Thus,

$$C = 16(36 + 21 + 12 + 8) + 36(21 + 8) + 11(12 + 8) + 36(8) = 2784.$$

Table 8.16 portrays this calculation of the total number of concordant pairs.

**TABLE 8.16:** Illustration of Calculation of Number of Concordant Pairs, *C*

| | NTH | PH | VH | NTH | PH | VH | NTH | PH | VH | NTH | PH | VH |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Below | 16 | | | | 36 | | | | | | | |
| Average | | 36 | 21 | | | 21 | 11 | | | | 36 | |
| Above | | 12 | 8 | | | 8 | | 12 | 8 | | | 8 |

$C = 16(36+21+12+8)$ $+ 36(21+8)$ $+ 11(12+8)$ $+ 36(8) = 2784$

To find the total number of discordant pairs *D*, start at the corner of the table that is the high level of one variable and the low level of the other. For example, the 15 subjects in the cell (below average, very satisfied) form discordant pairs when paired with the (11 + 36 + 2 + 12) subjects below and to the left in the table who are higher on income but lower on happiness. Multiply the count in each cell by the counts in all cells that are higher on income but lower in job satisfaction. The total number of discordant pairs is

$$D = 15(11 + 36 + 2 + 12) + 21(2 + 12) + 36(11 + 2) + 36(2) = 1749.$$

Table 8.17 portrays the calculation of the number of discordant pairs.

**TABLE 8.17:** Illustration of Calculation of Number of Discordant Pairs, *D*

| | NTH | PH | VH | NTH | PH | VH | NTH | PH | VH | NTH | PH | VH |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Below | | | 15 | | | | | 36 | | | | |
| Average | 11 | 36 | | | | 21 | 11 | | | | 36 | |
| Above | 2 | 12 | | 2 | 12 | | 2 | | | 2 | | |

$D = 15(11+36+2+12)$ $+ 21(2+12)$ $+ 36(11+2)$ $+ 36(2) = 1749$

In summary, Table 8.15 has $C = 2784$ and $D = 1749$. More pairs show evidence of a positive association (i.e., concordant pairs) than show evidence of a negative association (discordant pairs).

## Gamma

A positive difference for $C - D$ occurs when $C > D$. This indicates a positive association. A negative difference for $C - D$ reflects a negative association.

Larger sample sizes have larger numbers of pairs with, typically, larger absolute differences in $C - D$. Therefore, we standardize this difference to make it easier to interpret. To do this, we divide $C - D$ by the total number of pairs that are either concordant or discordant, $C + D$. This gives the measure of association called ***gamma***. Its sample formula is

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

Here are some properties of gamma:

- The value of gamma falls between $-1$ and $+1$.
- The sign of gamma indicates whether the association is positive or negative.
- The larger the absolute value of gamma, the stronger the association.

A table for which gamma equals 0.60 or $-0.60$ exhibits a stronger association than one for which gamma equals 0.30 or $-0.30$, for example. The value $+1$ represents the strongest positive association. This occurs when there are no discordant pairs ($D = 0$), so all the pairs reveal a positive association. Gamma equals $-1$ when $C = 0$, so all pairs reveal a negative association. Gamma equals 0 when $C = D$.

For Table 8.15, $C = 2784$ and $D = 1749$, so

$$\hat{\gamma} = \frac{2784 - 1749}{2784 + 1749} = 0.228.$$

This sample exhibits a positive association between family income and happiness. The higher the family income, the greater the happiness tends to be. However, the sample value is closer to 0 than to 1, so the association is relatively weak.

The calculation of gamma is rather messy. Most statistical software can find gamma for you.

### Gamma Is a Difference between Two Ordinal Proportions

Another interpretation for the magnitude of gamma follows from the expression

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

Now $(C + D)$ is the total number of pairs that are concordant or discordant. The ratio $C/(C + D)$ is the proportion of those pairs that are concordant, $D/(C + D)$ is the proportion of the pairs that are discordant, and $\hat{\gamma}$ is the difference between the two proportions.

For example, suppose $\hat{\gamma} = 0.60$. Then, since 0.80 and 0.20 are the two proportions that sum to 1 and have a difference of $0.80 - 0.20 = 0.60$, 80% of the pairs are concordant and 20% are discordant. Similarly, $\hat{\gamma} = -0.333$ indicates that 1/3 of the pairs are concordant and 2/3 of the pairs are discordant, since $1/3 + 2/3 = 1$ and $1/3 - 2/3 = -0.333$.

For Table 8.15, out of the $2784 + 1749 = 4533$ pairs that are concordant or discordant, the proportion $2784/4533 = 0.614$ are concordant and the proportion $1749/4533 = 0.386$ are discordant; $\hat{\gamma} = 0.228$ is the difference between these proportions.

### Common Properties of Ordinal Measures

Gamma is one of several ordinal measures of association. Others are ***Kendall's tau-b*** and ***tau-c***, ***Spearman's rho-b***, and ***rho-c***, and ***Somers' d***. All these measures are similar in their basic purposes and characteristics. For lack of space, we do not define these other measures, but we will list some common properties. These properties also hold for the *correlation* for quantitative variables, which was introduced in Section 3.5 and will be used extensively in the next chapter.

- Ordinal measures of association take values between $-1$ and $+1$. The sign tells us whether the association is positive or negative.
- If the variables are statistically independent, then the population values of ordinal measures equal 0.
- The stronger the association, the larger the absolute value of the measure. Values of 1.0 and $-1.0$ represent the strongest associations.
- With the exception of Somers' $d$, the ordinal measures of association named above do not distinguish between response and explanatory variables. They take the same value when variable $y$ is the response variable as when it is the explanatory variable.

So far, we have discussed the use of ordinal measures only for description. The next section presents statistical inference, namely, confidence intervals and tests for ordinal data.

## 8.6 INFERENCE FOR ORDINAL ASSOCIATIONS*

The chi-squared test of whether two categorical variables are independent treats the variables as nominal. Other tests are usually more powerful when the variables are ordinal. This section presents such a test and shows how to construct confidence intervals for ordinal measures of association such as gamma. The inferences are best applied to a large random sample. As a rough guideline, each of $C$ and $D$ should exceed about 50.

### Confidence Intervals for Measures of Association

Confidence intervals help us gauge the strength of the association in the population. Let $\gamma$ denote the population value of gamma. For sample gamma, $\hat{\gamma}$, its sampling distribution is approximately normal about $\gamma$. Its standard error $se$ describes the variation in $\hat{\gamma}$ values around $\gamma$ among samples of the given size. The formula for $se$ is complicated but it is reported by most software. A confidence interval for $\gamma$ has the form

$$\hat{\gamma} \pm z(se).$$

**EXAMPLE 8.8 Association between Income and Happiness**

For the data in Table 8.15 on family income and happiness, $\hat{\gamma} = 0.228$. We'll see in Table 8.18 that this has $se = 0.114$. A 95% confidence interval for $\gamma$ is

$$\hat{\gamma} \pm 1.96(se), \text{ or } 0.228 \pm 1.96(0.114), \text{ or } 0.228 \pm 0.223,$$

which equals (0.005, 0.45). We can be 95% confident that $\gamma$ is no less than 0.005 and no greater than 0.45. It is plausible that essentially no association exists between income and happiness, but it is also plausible that a moderate positive association exists. We need a larger sample size to estimate this more precisely. ■

### Test of Independence Using Gamma

Next we'll consider a test of independence that treats the variables as ordinal. As in the chi-squared test, the null hypothesis is that the variables are statistically independent. We express the test in terms of gamma, but a similar approach works with other ordinal measures of association. The alternative hypothesis can take the two-sided form $H_a$: $\gamma \neq 0$ or a one-sided form, $H_a$: $\gamma > 0$ or $H_a$: $\gamma < 0$, when we predict the direction of the association.

The test statistic has the $z$ statistic form. It takes the difference between $\hat{\gamma}$ and the value of 0 that gamma takes when $H_0$: independence is true and divides by the standard error,

$$z = \frac{\hat{\gamma} - 0}{se}.$$

This test statistic has approximately the standard normal distribution when $H_0$ is true. Some software also reports a $se$ and/or related $P$-value that holds only under $H_0$.

**EXAMPLE 8.9 Testing Independence between Income and Happiness**

Does Table 8.15 relating family income and happiness suggest these variables are associated in the population? The chi-squared test of independence has $\chi^2 = 3.82$ with $df = 4$, for which the $P$-value equals 0.43. This test does not show any evidence of an association. The chi-squared test treats the variables as nominal, however, and ordinal-level methods are more powerful if there is a positive or negative trend.

Table 8.18 shows a printout for the analysis of Table 8.15. The $\hat{\gamma} = 0.228$ value has $se = 0.114$, labelled as *Asymp. std. error*, where *Asymp.* stands for "asympotic" or "large-sample." The test statistic equals

$$z = \frac{\hat{\gamma} - 0}{se} = \frac{0.228 - 0}{0.114} = 2.00.$$

From the standard normal table, the $P$-value for $H_a$: $\gamma \neq 0$ equals 0.046. (SPSS reports a $P$-value of 0.050, based on using a different standard error in the test statistic that only applies under $H_0$.)

**TABLE 8.18:** Part of a Computer Printout for Analyzing Table 8.17

| | Value | DF | Asymp. Sig. |
|---|---|---|---|
| Pearson Chi-Square | 3.816 | 4 | 0.431 |

| | Value | Asymp. Std. Error | Approx. sig. |
|---|---|---|---|
| Gamma | 0.2283 | 0.1139 | 0.050 |

This test shows some evidence of an association. Since the sample value of gamma was positive, it seems that a positive association exists between income and happiness. The test for $H_a$: $\gamma > 0$ has $P = 0.023$ (or 0.025 using the null $se$). ■

### Ordinal Tests versus Pearson Chi-Squared Test

The $z$ test result for these data providing evidence of an association may seem surprising. The chi-squared statistic of $\chi^2 = 3.82$ with $df = 4$ provided no evidence ($P = 0.43$).

A test of independence based on an ordinal measure is usually preferred to the chi-squared test when both variables are ordinal. The $\chi^2$ statistic ignores the ordering of the categories, taking the same value no matter how the levels are ordered. If a positive or negative trend exists, ordinal measures are usually more powerful for detecting it. Unfortunately, the situation is not clear cut. It is possible for the chi-squared test to be more powerful even if the data are ordinal.

To explain this, we first note that the null hypothesis of independence is not equivalent to a value of 0 for population gamma. Although independence implies $\gamma = 0$, the converse is not true. Namely, $\gamma$ may equal 0 even though the variables are not statistically independent. For example, Table 8.19 shows a relationship between two variables that does not have a single trend. Over the first two columns there is a positive relationship, since $y$ increases when $x$ increases. Over the last two columns there is a negative relationship, as $y$ decreases when $x$ increases. For the entire table, $C = 25(25 + 25) = 1250 = D$, so $\gamma = 0$. The proportion of concordant pairs equals the proportion of discordant pairs. However, there is not independence, because the conditional distribution on $y$ for the low level of $x$ is completely different from the conditional distribution on $y$ for the high level of $x$.

**TABLE 8.19:** A Relationship for Which Ordinal Measures of Association Equal 0. The variables are dependent even though gamma equals 0.

| | | Level of $y$ | | | |
|---|---|---|---|---|---|
| | | Very low | Low | High | Very high |
| Level of $x$ | Low | 25 | 0 | 0 | 25 |
| | High | 0 | 25 | 25 | 0 |

Thus, an ordinal measure of association may equal 0 when the variables are statistically dependent but the dependence does not have an overall positive or overall negative trend. The chi-squared test can perform better than the ordinal test when the relationship does not have a single trend. In practice, most relationships with ordinal variables have primarily one trend, if any. So the ordinal test is usually more powerful than the chi-squared test.

### Similar Inference Methods for Other Ordinal Measures

The inference methods for gamma apply also to other ordinal measures of association. For a confidence interval, take the sample value and add and subtract a $z$-score times the standard error, which is available using software. Test results are usually similar for any ordinal measure based on the difference between the numbers of concordant pairs and discordant pairs, such as gamma or Kendall's tau-$b$.

An alternative approach to detect trends assigns scores to the categories for each variable and uses the correlation and a $z$ test based on it. (Section 9.5 presents a closely related test.) Some software reports this as a test of *linear-by-linear association*.

Whenever possible, it is better to choose the categories for ordinal variables finely rather than crudely. For instance, it is better to use four or five categories than only two categories. Standard errors of measures tend to be smaller with more categories, for a given sample size. Thus, the finer the categorizations, the shorter the confidence interval for a population measure of association tends to be. In addition, finer measurement makes it more valid to treat the data as quantitative and use the more powerful methods presented in the following chapter for quantitative variables.

### Mixed Ordinal–Nominal Contingency Tables

For a cross-classification of an ordinal variable with a nominal variable that has only two categories, ordinal measures of association are still valid. In that case, the sign of the measure indicates which level of the nominal variable is associated with higher responses on the ordinal variable. For instance, suppose gamma $= -0.12$ for the association in a $2 \times 3$ table relating gender (female, male) to happiness (not too happy, pretty happy, very happy). Since the sign is negative, the "higher" level of gender (i.e., male) tends to occur with lower happiness. The association is weak, however.

When the nominal variable has more than two categories, it is inappropriate to use an ordinal measure such as gamma. There are specialized methods for mixed nominal–ordinal tables, but it is usually simplest to treat the ordinal variable as quantitative by assigning scores to its levels. The methods of Chapter 12, which generalize comparisons of two means to several groups, are then appropriate. Section 15.4 presents a modeling approach that does not require assigning scores to ordinal response variables.

## 8.7 CHAPTER SUMMARY

This chapter introduced analyses of association for categorical variables:

- By *describing the counts* in ***contingency tables*** using percentage distributions, called ***conditional distributions***, across the categories of the response variable. If the population conditional distributions are identical, the two variables are ***statistically independent***—the probability of any particular response is the same for each level of the explanatory variable.
- By using ***chi-squared*** to ***test*** $H_0$***: independence*** between the variables. The $\chi^2$ test statistic compares each observed frequency $f_o$ to the expected frequency $f_e$ satisfying $H_0$, using

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}.$$

  The test statistic has a large-sample chi-squared distribution. The ***degrees of freedom*** depend on the number of rows $r$ and the number of columns $c$, through $df = (r - 1)(c - 1)$. The $P$-value is the right-tail probability above the observed value of $\chi^2$.
- By *describing the pattern of association* using ***standardized residuals*** for the cells in the table. A standardized residual reports the number of standard errors that $(f_o - f_e)$ falls from 0. A value larger than about 2 or 3 in absolute value indicates that that cell provides evidence of association in a particular direction.
- By *describing the strength of association*. For $2 \times 2$ tables the ***difference of proportions*** is useful, as is the ***odds ratio***, the ratio of odds from the two rows. Each odds measures the proportion of successes divided by the proportion of failures. When there is independence, the difference of proportions equals 0 and the odds ratio equals 1. The stronger the association, the farther the measures fall from these baseline values.

This chapter also presented methods for analyzing association between two ordinal variables.

- Many ***ordinal measures of association*** use the numbers of ***concordant pairs*** (the subject who is higher on $x$ also is higher on $y$) and ***discordant pairs*** (the subject who is higher on $x$ is lower on $y$).
- Of the pairs that are concordant or discordant, ***gamma*** equals the difference between the proportions of the two types. Gamma falls between $-1$ and $+1$, with larger absolute values indicating stronger association. When the variables are independent, gamma equals 0.

The chi-squared test treats the data as nominal. When the variables are ordinal, methods that use the ordinality (such as a $z$ test based on sample gamma) are more powerful for detecting a positive or negative association trend.

The next chapter introduces similar methods for describing and making inferences about the association between two quantitative variables.

## PROBLEMS

### Practicing the Basics

**8.1.** GSS surveys routinely show that in the United States, about 40% of males and 40% of females believe that a women should be able to get an abortion if she wants it for any reason (variable ABANY).

**(a)** Construct a contingency table showing the conditional distribution on whether unrestricted abortion should be legal (yes, no) by gender.

**(b)** Based on these results, does statistical independence seem plausible between gender and opinion about unrestricted abortion? Explain.

**8.2.** Whether a woman becomes pregnant in the next year is a categorical variable with categories (yes, no), and whether she and her partner use contraceptives is another categorical variable with categories (yes, no). Would you expect these variables to be statistically independent, or associated? Explain.

**8.3.** Every year, a large-scale poll of college freshmen conducted by the Higher Education Research Institute at UCLA asks their opinions about a variety of issues. In 2002, 46% of men and 35% of women in the survey of 283,000 college freshmen indicated support for legalization of marijuana.

**(a)** If results for the population of college freshmen were similar to these, would gender and opinion about legalizing marijuana be independent, or dependent?

**(b)** Display hypothetical population percentages in a contingency table for which these variables would be independent.

**8.4.** Some political analysts claimed that during the presidency of George W. Bush, the popularity of the U.S. decreased dramatically around the world. In *America Against the World: How We Are Different and Why We Are Disliked*,[1] the Pew Research Center summarized results of 91,000 interviews conducted in 51 nations. In Germany, for example, the study reported that those having favorable opinions of the U.S. changed between 2000 and 2006 from 78% to 37%. Show how to construct a contingency table relating opinion about the U.S. by year of survey, for Germany. For this table, identify the response variable, the explanatory variable, and the conditional distributions.

**8.5.** Based on current estimates of how well mammograms detect breast cancer, Table 8.20 shows what to expect for 100,000 adult women over the age of 40 in terms of whether a woman has breast cancer and whether a mammogram gives a positive result (i.e., indicates that the woman has breast cancer).

**(a)** Construct the conditional distributions for the mammogram test result, given the true disease status. Does the mammogram appear to be a good diagnostic tool?

**(b)** Construct the conditional distribution of disease status, for those who have a positive test result. Use this to explain why even a good diagnostic test can have a high false positive rate when a disease is not common.

**TABLE 8.20**

| | | Diagnostic Test | |
|---|---|---|---|
| | | Positive | Negative |
| Breast Cancer | Yes | 860 | 140 |
| | No | 11,800 | 87,120 |

**8.6.** Data posted at the FBI Web site (www.fbi.gov) indicated that of all blacks slain in 2005, 91% were slain by blacks, and of all whites slain in 2005, 83%

[1]Kohut, A. and Stokes, B. (2006). *America Against the World: How We Are Different and Why We Are Disliked*. Times Books.

were slain by whites. Let $y$ denote race of victim and $x$ denote race of murderer.

**(a)** Which conditional distributions do these statistics refer to, those of $y$ at given levels of $x$ or those of $x$ at given levels of $y$? Set up a contingency table showing these distributions.

**(b)** Are $x$ and $y$ independent or dependent? Explain.

**8.7.** How large a $\chi^2$ value provides a $P$-value of 0.05 for testing independence for the following table dimensions?

**a)** $2 \times 2$ **b)** $3 \times 3$ **c)** $2 \times 5$ **d)** $5 \times 5$ **e)** $3 \times 9$

**8.8.** Show that the contingency table in Table 8.21 has four degrees of freedom by showing how the four cell counts given determine the others.

**TABLE 8.21**

| | | | |
|---|---|---|---|
| 10 | 20 | | 60 |
| 30 | 40 | | 100 |
| | | | 40 |
| 50 | 80 | 70 | |

**8.9.** In 2000 the GSS asked whether a subject is willing to accept cuts in the standard of living to help the environment (GRNSOL), with categories (very willing, fairly willing, neither willing nor unwilling, not very willing, not at all willing). When this was cross-tabulated with sex, $\chi^2 = 8.0$.

**(a)** What are the hypotheses for the test to which refers?

**(b)** Report the $df$ value on which $\chi^2$ is based.

**(c)** What conclusion would you make, using a significance level of (i) 0.05, (ii) 0.10? State your conclusion in the context of this study.

**8.10.** Table 8.22 refers to a survey of senior high school students in Dayton, Ohio.

**(a)** Construct conditional distributions that treat cigarette smoking as the response variable. Interpret.

**(b)** Test whether cigarette use and alcohol use are statistically independent. Report the $P$-value and interpret.

**TABLE 8.22**

| | | Cigarette Use | |
|---|---|---|---|
| | | Yes | No |
| Alcohol Use | Yes | 1449 | 500 |
| | No | 46 | 281 |

*Source*: Thanks to Professor Harry Khamis for providing these data.

**8.11.** Are people happier who believe in life after death? Go to the GSS Web site sda.berkeley.edu/GSS and download the contingency table for the 2006 survey relating happiness and whether you believe in life after death (variables HAPPY and POSTLIFE, with YEAR(2006) in the 'selection filter').

**(a)** State a research question that could be addressed with the output.

**(b)** Report the conditional distributions, using happiness as the response variable, and interpret.

**(c)** Report the $\chi^2$ value and its $P$-value. (You can get this by checking 'Statistics'.) Interpret.

**(d)** Interpret the standardized residuals. (You can get them by checking 'z-statistic'.)

**8.12.** In the GSS, subjects who were married were asked the happiness of their marriage, the variable coded as HAPMAR.

**(a)** Go to sda.berkeley.edu/GSS/ and construct a contingency table for 2006 relating HAPMAR to family income measured as (above average, average, below average), by entering FINRELA(r: 1-2; 3; 4-5) as the row variable and YEAR(2006) in the selection filter. Use a table or graph with conditional distributions to describe the association.

**(b)** By checking 'Statistics,' you request the chi-squared statistic. Report it and its $df$ and $P$-value, and interpret.

**8.13.** The sample in Table 8.15 is 157 black Americans. Table 8.23 shows cell counts and standardized residuals for income and happiness for white subjects in the 2004 GSS.

**(a)** Explain how to interpret the Pearson chi-squared statistic and its associated $P$-value.

**(b)** Explain how to interpret the standardized residuals in the four corner cells.

**TABLE 8.23**

Rows: income Columns: happiness

| | not | pretty | very | All |
|---|---|---|---|---|
| below | 62 | 187 | 45 | 294 |
| | 5.34 | 3.43 | -7.40 | |
| average | 47 | 270 | 181 | 498 |
| | -2.73 | -0.57 | 2.53 | |
| above | 22 | 127 | 118 | 267 |
| | -2.37 | -2.88 | 4.73 | |
| All | 131 | 584 | 131 | 1059 |

Cell Contents: Count
Standardized residual

Pearson Chi-Square = 72.15, DF = 4, P-Value = 0.000

**8.14.** Table 8.24 shows SPSS analyses with the 2004 GSS, for variables party ID and race.

**(a)** Report the expected frequency for the first cell, and show how SPSS obtained it.

**(b)** Test the hypothesis of independence between party ID and race. Report the test statistic and $P$-value and interpret.

**(c)** Use the standardized residuals (labelled ADJ RES here for "adjusted residuals" to describe the pattern of association.

**TABLE 8.24**

```
            Count   PARTY_ID
            Exp Val
            Adj Res                              Row
                      democr   indep    repub   Total
RACE        black       250     106       17     373
                      129.1   129.0    114.9
                       14.2    -2.7    -11.9

            white       640     783     1775    2198
                      760.9   760.0    677.1
                      -14.2     2.7     11.9

       Column Total     890     889      792    2571

   Chi-Square           Value      DF     Significance
------------------    ---------   ----   ------------
Pearson                 234.73      2       0.0000
```

**8.15.** For a 2 × 4 cross classification of gender and religiosity (very, moderately, slightly, not at all) for recent GSS data, the standardized residual was 3.2 for females who are very religious, −3.2 for males who are very religious, −3.5 for females who are not at all religious, and 3.5 for males who are not at all religious. All other standardized residuals fell between −1.1 and 1.1. Interpret.

**8.16.** Table 8.25 is from the 2006 General Social Survey, cross-classifying happiness (HAPPY) and marital status (MARITAL).

**TABLE 8.25**

| Marital Status | Very Happy | Pretty Happy | Not Too Happy |
|---|---|---|---|
| Married | 600 (13.1) | 720 (−5.4) | 93 (−10.0) |
| Widowed | 63 (−2.2) | 142 (−0.2) | 51 (3.4) |
| Divorced | 93 (−6.1) | 304 (3.2) | 88 (3.6) |
| Separated | 19 (−2.7) | 51 (−1.2) | 31 (5.3) |
| Never Married | 144 (−7.4) | 459 (4.2) | 127 (4.0) |

**(a)** Software reports that $\chi^2 = 236.4$. Interpret.

**(b)** Table 8.25 also shows, in parentheses, the standardized residuals. Summarize the association by indicating which marital statuses have strong evidence of (i) more, (ii) fewer people in the population in the *very happy* category than if the variables were independent.

**(c)** Compare the married and divorced groups by the difference in proportions in the *very happy* category.

**8.17.** In a *USA Today*/Gallup poll in July 2006, 82% of Republicans approved of President George W. Bush's performance, whereas 9% of Democrats approved. Would you characterize the association between political party affiliation and opinion about Bush's performance as weak, or strong? Explain why.

**8.18.** In a recent GSS, the death penalty for subjects convicted of murder was favored by 74% of whites and 43% of blacks. It was favored by 75% of males and 63% of females. In this sample, which variable was more strongly associated with death penalty opinion—race or gender? Explain why.

**8.19.** Refer to Exercise 8.10, on alcohol use and cigarette use.

**(a)** Describe the strength of association using the difference between users and nonusers of alcohol in the proportions who have used cigarettes. Interpret.

**(b)** Describe the strength of association using the difference between users and nonusers of cigarettes in the proportions who have used alcohol. Interpret.

**(c)** Describe the strength of association using the odds ratio. Interpret. Does the odds ratio value depend on your choice of response variable?

**8.20.** Table 8.26 cross-classifies 68,694 passengers in autos and light trucks involved in accidents in the state of Maine by whether they were wearing a seat belt and by whether they were injured or killed. Describe the association using

**(a)** The difference between two proportions, treating whether injured or killed as the response variable.

**(b)** The odds ratio.

**TABLE 8.26**

| | | Injury | |
|---|---|---|---|
| | | Yes | No |
| Seat Belt | Yes | 2409 | 35,383 |
| | No | 3865 | 27,037 |

*Source*: Thanks to Dr. Cristanna Cook, Medical Care Development, Augusta, Maine, for supplying these data.

**8.21.** According to the Substance Abuse and Mental Health Archive, a 2003 national household survey on drug abuse indicated that for Americans aged

26–34, 51% had used marijuana at least once in their lifetime, and 18% had used cocaine at least once.

**(a)** Find the odds of having used (i) marijuana, (ii) cocaine. Interpret.

**(b)** Find the odds ratio comparing marijuana use to cocaine use Interpret.

**8.22.** According to the U.S. Department of Justice, in 2004 the incarceration rate in the nation's prisons was 1 per 109 male residents, 1 per 1563 female residents, 1694 per 100,000 black residents, and 252 per 100,000 white residents (*Source*: www.ojp.usdoj.gov/bjs).

**(a)** Find the odds ratio between whether incarcerated and (i) gender, (ii) race. Interpret.

**(b)** According to the odds ratio, which has the stronger association with whether incarcerated, gender or race? Explain.

**8.23.** Refer to Table 8.1 (page 222) on political party ID and gender. Find and interpret the odds ratio for each 2 × 2 subtable. Explain why this analysis suggests that the last two columns show essentially no association.

**8.24.** For college freshmen in 2004, the percent who agreed that homosexual relationships should be legally prohibited was 38.0% of males and 23.4% of females (www.gseis.ucla.edu/heri/american_freshman.html).

**(a)** The odds ratio is 2.01. Explain what is wrong with the interpretation, "The probability of a yes response for males is 2.01 times the probability of a yes response for females." Give the correct interpretation.

**(b)** The odds of a yes response equaled 0.613 for males. Estimate the probability of a yes response for males.

**(c)** Based on the odds of 0.613 for males and the odds ratio of 2.01, show how to estimate the probability of a yes response for females.

**8.25.** Table 8.27 cross-classifies happiness with family income for the subsample of the 2004 GSS that identified themselves as Jewish.

**(a)** Find the number of (i) concordant pairs, (ii) discordant pairs.

**(b)** Find gamma and interpret.

**TABLE 8.27**

| | | HAPPY | | |
|---|---|---|---|---|
| | | Not_too | Pretty | Very |
| INCOME | Below | 1 | 2 | 1 |
| | Average | 0 | 5 | 2 |
| | Above | 2 | 4 | 0 |

**(c)** Show how to express gamma as a difference between two proportions.

**8.26.** For the 2006 GSS, $\hat{\gamma} = 0.22$ for the relationship between job satisfaction (SATJOB; categories very dissatisfied, little dissatisfied, moderately satisfied, very satisfied) and family income (FINRELA; below average, average, above average).

**(a)** Would you consider this a very strong or relatively weak association? Explain.

**(b)** Of the pairs that are concordant or discordant, what proportion are concordant? Discordant?

**(c)** Is this a stronger or a weaker association than the one between job satisfaction and happiness (variable HAPPY), which has $\hat{\gamma} = 0.40$? Explain.

**8.27.** A study on educational aspirations of high school students[2] measured aspirations using the scale (some high school, high school graduate, some college, college graduate) and family income with three ordered categories. Software provides the results shown in Table 8.28.

**(a)** Use gamma to summarize the association.

**(b)** Test independence of educational aspirations and family income using the chi-squared test. Interpret.

**(c)** Find the 95% confidence interval for gamma. Interpret.

**(d)** Conduct an alternative test of independence that takes category ordering into account. Why are results so different from the chi-squared test?

**TABLE 8.28**

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 6 | 8.871 | 0.181 |

| Statistic | Value | ASE |
|---|---|---|
| Gamma | 0.163 | 0.080 |

**8.28.** Refer to Exercise 8.13, on happiness and income. The analysis there does not take into account the ordinality of the variables. Using software:

**(a)** Summarize the strength of association by finding and interpreting gamma.

**(b)** Construct and interpret a 95% confidence interval for the population value of gamma.

## Concepts and Applications

**8.29.** Refer to the "Student survey" data file (Exercise 1.11 on page 8). Using software, create and

[2] S. Crysdale, *Intern. J. Compar. Sociol.*, vol. 16, 1975, pp. 19–36.

analyze descriptively and inferentially the contingency table relating opinion about abortion and **(a)** political affiliation, **(b)** religiosity.

**8.30.** Refer to the data file you created in Exercise 1.12. For variables chosen by your instructor, pose a research question and conduct descriptive and inferential statistical analyses. Interpret and summarize your findings in a short report.

**8.31.** In 2002 the GSS asked how housework was shared between the respondent and his or her spouse (HHWKFAIR). Possible responses were 1 = I do much more than my fair share, 2 = I do a bit more than my fair share, 3 = I do roughly my fair share, 4 = I do a bit less than my fair share, 5 = I do much less than my fair share. Table 8.29 shows results according to the respondent's sex. State a research question that could be addressed with this output, and prepare a one-page report summarizing what you learn. (The "Adj. Residual" is the standardized residual.)

**8.32.** Pose a research question about attitude regarding homosexual relations and political ideology. Using the most recent GSS data on HOMOSEX and POLVIEWS, conduct a descriptive and inferential analysis to address this question. Prepare a short report summarizing your analysis.

**8.33.** Several sociologists have reported that racial prejudice varies according to religious group. Examine this using Table 8.30, for white respondents to the 2002 GSS. The variables are Fundamentalism/Liberalism of Respondent's Religion (FUND) and response to the question (RACMAR), "Do you think there should be laws against marriages between blacks and whites?" Analyze these data. Prepare a report, describing your analyses and providing interpretations of the data.

**TABLE 8.30**

| Religious Preference | Laws against Marriage | | Total |
|---|---|---|---|
| | Favor | Oppose | |
| Fundamentalist | 39 | 142 | 181 |
| Moderate | 21 | 248 | 269 |
| Liberal | 17 | 236 | 253 |
| None | 16 | 74 | 90 |
| Total | 93 | 700 | 793 |

**8.34.** For 2006 GSS data, of those identifying as Democrats, 616 classified themselves as liberal and 262 as conservative. Of those identifying as Republicans, 94 called themselves liberal and 721 called themselves conservative. Using methods presented in this chapter, describe the strength of association.

**8.35.** A study[3] of American armed forces who had served in Iraq or Afghanistan found that the event of being attacked or ambushed was reported by 1139 of 1961 Army members who had served in Afghanistan, 789 of 883 Army members who had served in Iraq, and 764 of 805 Marines who had served in Iraq. Summarize these data using conditional distributions and measures of association.

**8.36.** Shortly before a gubernatorial election, a poll asks a random sample of 50 potential voters the following questions:

> Do you consider yourself to be a Democrat (D), a Republican (R), or Independent (I)?
>
> If you were to vote today, would you vote for the Democratic candidate (D), the

**TABLE 8.29**

| | | | HHWKFAIR 1 | 2 | 3 | 4 | 5 | Total |
|---|---|---|---|---|---|---|---|---|
| sex | female | Count | 121 | 108 | 135 | 19 | 6 | 389 |
| | | % within sex | 31.1% | 27.8% | 34.7% | 4.9% | 1.5% | 100.0% |
| | | Adj. Residual | 8.0 | 5.9 | -4.2 | -7.1 | -4.9 | |
| | male | Count | 18 | 28 | 148 | 68 | 29 | 291 |
| | | % within sex | 6.2% | 9.6% | 50.9% | 23.4% | 10.0% | 100.0% |
| | | Adj. Residual | -8.0 | -5.9 | 4.2 | 7.1 | 4.9 | |

| | Value | df | Asymp. Sig. |
|---|---|---|---|
| Pearson Chi-Square | 155.8 | 4 | .000 |

| | Value | Asymp. Std. Error |
|---|---|---|
| Gamma | .690 | .038 |

[3] C. Hoge et al., *New England J. Medic.*, vol. 351, 2004, pp. 13–21.

Republican (R), or would you be undecided (U) about how to vote?

Do you plan on voting in the election? Yes (Y) or no (N)?

For each person interviewed, the answers to the three questions are entered in a data file. For example, the entry (D, U, N) represents a registered Democrat who is undecided and who does not expect to vote. Table 8.31 summarizes results of the 50 interviews. Using software, create a data file and conduct the following analyses:

**(a)** Construct the 3 × 3 contingency table relating party affiliation to intended vote. Report the conditional distributions on intended vote for each of the three party affiliations. Are they very different?

**(b)** Report the result of the test of the hypothesis that intended vote is independent of party affiliation. Provide the test statistic and the $P$-value, and interpret the result.

**(c)** Supplement the analyses in (a)–(b) to investigate the association more fully. Interpret.

**TABLE 8.31**

| | | | | |
|---|---|---|---|---|
| (D, U, N) | (R, R, Y) | (I, D, Y) | (I, U, N) | (R, U, N) |
| (I, D, N) | (R, R, Y) | (I, U, N) | (D, U, Y) | (D, R, N) |
| (I, D, N) | (D, D, Y) | (D, D, Y) | (I, D, Y) | (R, U, N) |
| (D, R, N) | (R, D, N) | (D, U, N) | (D, D, Y) | (R, R, Y) |
| (R, R, Y) | (D, D, N) | (D, D, Y) | (I, D, Y) | (R, R, N) |
| (D, D, Y) | (D, R, Y) | (I, U, N) | (D, D, N) | (D, D, Y) |
| (R, R, Y) | (R, R, Y) | (D, U, N) | (I, R, N) | (I, R, Y) |
| (R, R, Y) | (I, U, Y) | (D, D, Y) | (D, R, Y) | (D, D, N) |
| (D, D, Y) | (I, R, Y) | (R, R, Y) | (I, D, Y) | (R, R, N) |
| (R, R, Y) | (D, D, Y) | (I, D, Y) | (I, R, N) | (R, R, Y) |

**8.37.** **(a)** When the sample size is very large, we have not necessarily established an important result when we show a statistically significant association. Explain.

**(b)** The remarks in Sections 8.3 and 8.4 about small $P$-values not necessarily referring to an important effect apply for any significance test. Explain why, discussing the effect of $n$ on standard errors and the sizes of test statistics.

**8.38.** Answer true or false for the following. Explain your answer.

**(a)** Even when the sample conditional distributions in a contingency table are only slightly different, when the sample size is very large it is possible to have a large $\chi^2$ test statistic and a very small $P$-value for testing $H_0$: independence.

**(b)** If the odds ratio = 2.0 between gender (female, male) and opinion on some issue (favor, oppose), then the odds ratio = −2.0 if we measure gender as (male, female).

**(c)** Interchanging two rows in a contingency table has no effect on the chi-squared statistic.

**(d)** Interchanging two rows in a contingency table has no effect on gamma.

**(e)** If $\gamma = 0$ for two variables, then the variables are statistically independent.

**8.39.** The correct answer in Exercise 8.38(c) implies that if the chi-squared statistic is used for a contingency table having ordered categories in both directions, then (select the correct response(s))

**(a)** The statistic actually treats the variables as nominal.

**(b)** Information about the ordering is ignored.

**(c)** The test is usually not as powerful for detecting association as a test statistic based on numbers of concordant and discordant pairs.

**(d)** The statistic cannot differentiate between positive and negative associations.

**8.40.** Each subject in a sample of 100 men and 100 women is asked to indicate which of the following factors (one or more) are responsible for increases in crime committed by teenagers: A—the increasing gap in income between the rich and poor, B—the increase in the percentage of single-parent families, C—insufficient time that parents spend with their children, D—criminal penalties given by courts are too lenient, E—increasing problems with drugs in society, F—increasing levels of violence shown on TV. To analyze whether responses differ by gender of respondent, we cross-classify the responses by gender, as Table 8.32 shows.

**(a)** Is it valid to apply the chi-squared test of independence to these data? Explain.

**(b)** Explain how this table actually provides information needed to cross-classify gender with each of six variables. Construct the contingency table relating gender to opinion about whether the increasing gap in income is responsible for increases in teenage crime.

**TABLE 8.32**

| Gender | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| Men | 60 | 81 | 75 | 63 | 86 | 62 |
| Women | 75 | 87 | 86 | 46 | 82 | 83 |

***8.41.** Table 8.33 exhibits the maximum possible association between two binary variables for a sample of size $n$.

**(a)** Show that $\chi^2 = n$ for this table and, hence, that the maximum value of $\chi^2$ for 2 × 2 tables is $n$.

**(b)** The ***phi-squared*** measure of association for 2 × 2 contingency tables has sample value

$$\hat{\phi}^2 = \frac{\chi^2}{n}.$$

Explain why this measure falls between 0 and 1, with a population value of 0 corresponding to independence. (It is a special case, for 2 × 2 tables, of the *Goodman and Kruskal tau* measure and of the $r^2$ measure introduced in the next chapter.)

TABLE 8.33

| | |
|---|---|
| $n/2$ | 0 |
| 0 | $n/2$ |

***8.42.** For 2 × 2 tables, gamma simplifies to a measure first proposed about 1900 by the statistician G. Udny Yule, who also introduced the odds ratio. In that special case, gamma is called ***Yule's Q***.

**(a)** Show that for a generic table with counts $(a, b)$ in row 1 and $(c, d)$ in row 2, the number of concordant pairs equals $ad$, the number of discordant pairs equals $bc$, and $Q = (ad - bc)/(ad + bc)$.

**(b)** Show that the absolute value of gamma equals 1 for any 2 × 2 table in which one of the cell frequencies is 0.

***8.43.** Construct a 3 × 3 table for each of the following conditions:

**(a)** Gamma equals 1. (*Hint*: There should be no discordant pairs.)

**(b)** Gamma equals −1.

**(c)** Gamma equals 0.

***8.44.** A chi-squared variable with degrees of freedom equal to $df$ has representation $z_1^2 + \ldots + z_{df}^2$, where $z_1, \ldots, z_{df}$ are independent standard normal variates.

**(a)** If $z$ is a test statistic that has a standard normal distribution, what distribution does $z^2$ have?

**(b)** Explain how to get the chi-squared values for $df = 1$ in Table C from $z$-scores in the standard normal table (Table A). Illustrate for the chi-squared value of 6.63 having $P$-value 0.01.

**(c)** The chi-squared statistic for testing $H_0$: independence between belief in an afterlife (yes, no) and happiness (not too happy, pretty happy, very happy) is $\chi_1^2$ in a 2 × 3 table for men and $\chi_2^2$ in a 2 × 3 table for women. If $H_0$ is true for each gender, then what is the probability distribution of $\chi_1^2 + \chi_2^2$?

***8.45.** For a 2 × 2 table with cell counts $a, b, c, d$, the sample log odds ratio $\log \hat{\theta}$ has approximately a normal sampling distribution with estimated standard error

$$se = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}.$$

The antilogs of the endpoints of the confidence interval for $\log(\theta)$ are endpoints of the confidence interval for $\theta$. For Table 8.13 on page 236, show that $se = 0.0833$ and the 95% confidence interval for the odds ratio is (67.3, 93.2). Interpret.