CHAPTER

17

The Chi-Square Statistic: Tests for Goodness of Fit and Independence

Tools You Will Need

The following items are considered essential background material for this chapter. If you doubt your knowledge of any of these items, you should review the appropriate chapter or section before proceeding.

- Proportions (math review, Appendix A)
- Frequency distributions (Chapter 2)

Preview

Loftus and Palmer (1974) conducted a classic experiment demonstrating how language can influence eyewitness memory. A sample of 150 students watched a film of an automobile accident. After watching the film, the students were separated into three groups. One group was asked, "About how fast were the cars going when they smashed into each other?" Another group received the same question except that the verb was changed to "hit" instead of "smashed into." A third group served as a control and was not asked any question about the speed of the two cars. A week later, the participants returned and were asked if they remembered seeing any broken glass in the accident. (There was no broken glass in the film.) Notice that the researchers are manipulating the form of the initial question and then measuring a yes/no response to a follow-up question 1 week later. Table 17.1 shows the structure of this design represented by a matrix with the independent variable (different groups) determining the rows of the matrix and the two categories for the dependent variable (yes/no) determining the columns. The number in each cell of the matrix is the frequency count showing how many participants are classified in that category. For example, of the 50 students who heard the word *smashed,* there were 16 (32%) who claimed to remember seeing broken glass even though there was none in the film. By comparison, only 7 of the 50 students (14%) who heard the word *hit* said they recalled seeing broken glass. The researchers would like to use these data to support the argument that a witness's "memory" can be influenced by the language used during questioning. If the two cars *smashed* into each other, then there must have been some broken glass.

The Problem: Although the Loftus and Palmer study involves an independent variable (the form of the question) and a dependent variable (memory for broken glass), you should realize that this study is different from any experiment we have considered in the past. Specifically, the Loftus and Palmer study does not produce a numerical score for each participant. Instead, each participant is simply classified into one of two categories (yes or no). The data consist of *frequencies* or *proportions* describing how many individuals are in each category. You should also note that Loftus and Palmer want to use a hypothesis test to evaluate the data. The null hypothesis would state that the form of the question has no effect on the memory of the witness. The hypothesis test would determine whether the sample data provide enough evidence to reject this null hypothesis.

Because there are no numerical scores, it is impossible to compute a mean or a variance for the sample data. Therefore, it is impossible to use any of the familiar hypothesis tests (such as a *t* test or analysis of variance [ANOVA]) to determine whether there is a significant difference between the treatment conditions. What is needed is a new hypothesis testing procedure that can be used with non-numerical data.

The Solution: In this chapter we introduce two hypothesis tests based on the *chi-square* statistic. Unlike earlier tests that require numerical scores (*X* values), the chi-square tests use sample frequencies and proportions to test hypothesis about the corresponding population values.

TABLE 17.1

A frequency distribution table showing the number of participants who answered either yes or no when asked whether they recalled seeing any broken glass 1 week after witnessing an automobile accident. Immediately after the accident, one group was asked how fast the cars were going when they smashed into each other. A second group was asked how fast the cars were going when they hit each other. A third group served as a control and was not asked about the speed of the cars.

Verb Used to Ask about the Speed of the Cars	Response to the Question: Did You See Any Broken Glass? Yes	No
Smashed into	16	34
Hit	7	43
Control (Not Asked)	6	44

17.1 PARAMETRIC AND NONPARAMETRIC STATISTICAL TESTS

All of the statistical tests that we have examined thus far are designed to test hypotheses about specific population parameters. For example, we used *t* tests to assess hypotheses about a population mean (μ) or mean difference ($\mu_1 - \mu_2$). In addition, these tests typically make assumptions about other population parameters. Recall that, for analysis of variance (ANOVA), the population distributions are assumed to be normal and homogeneity of variance is required. Because these tests all concern parameters and require assumptions about parameters, they are called *parametric tests*.

Another general characteristic of parametric tests is that they require a numerical score for each individual in the sample. The scores then are added, squared, averaged, and otherwise manipulated using basic arithmetic. In terms of measurement scales, parametric tests require data from an interval or a ratio scale (see Chapter 1).

Often, researchers are confronted with experimental situations that do not conform to the requirements of parametric tests. In these situations, it may not be appropriate to use a parametric test. Remember that when the assumptions of a test are violated, the test may lead to an erroneous interpretation of the data. Fortunately, there are several hypothesis-testing techniques that provide alternatives to parametric tests. These alternatives are called *nonparametric tests*.

In this chapter, we introduce two commonly used examples of nonparametric tests. Both tests are based on a statistic known as chi-square and both tests use sample data to evaluate hypotheses about the proportions or relationships that exist within populations. Note that the two chi-square tests, like most nonparametric tests, do not state hypotheses in terms of a specific parameter and they make few (if any) assumptions about the population distribution. For the latter reason, nonparametric tests sometimes are called *distribution-free tests*.

One of the most obvious differences between parametric and nonparametric tests is the type of data they use. All of the parametric tests that we have examined so far require numerical scores. For nonparametric tests, on the other hand, the participants are usually just classified into categories such as Democrat and Republican, or High, Medium, and Low IQ. Note that these classifications involve measurement on nominal or ordinal scales, and they do not produce numerical values that can be used to calculate means and variances. Instead, the data for many nonparametric tests are simply frequencies—for example, the number of Democrats and the number of Republicans in a sample of $n = 100$ registered voters.

Occasionally, you have a choice between using a parametric and a nonparametric test. Changing to a nonparametric test usually involves transforming the data from numerical scores to nonnumerical categories. For example, you could start with numerical scores measuring self-esteem and create three categories consisting of high, medium, and low self-esteem. In most situations, the parametric test is preferred because it is more likely to detect a real difference or a real relationship. However, there are situations for which transforming scores into categories might be a better choice.

1. Occasionally, it is simpler to obtain category measurements. For example, it is easier to classify students as high, medium, or low in leadership ability than to obtain a numerical score measuring each student's ability.
2. The original scores may violate some of the basic assumptions that underlie certain statistical procedures. For example, the *t* tests and ANOVA assume that the data come from normal distributions. Also, the independent-measures

tests assume that the different populations all have the same variance (the homogeneity-of-variance assumption). If a researcher suspects that the data do not satisfy these assumptions, it may be safer to transform the scores into categories and use a nonparametric test to evaluate the data.

3. The original scores may have unusually high variance. Variance is a major component of the standard error in the denominator of t statistics and the error term in the denominator of F-ratios. Thus, large variance can greatly reduce the likelihood that these parametric tests will find significant differences. Converting the scores to categories essentially eliminates the variance. For example, all individuals fit into three categories (high, medium, and low), no matter how variable the original scores are.
4. Occasionally, an experiment produces an undetermined, or infinite, score. For example, a rat may show no sign of solving a particular maze after hundreds of trials. This animal has an infinite, or undetermined, score. Although there is no absolute number that can be assigned, you can say that this rat is in the highest category, and then classify the other scores according to their numerical values.

17.2 THE CHI-SQUARE TEST FOR GOODNESS OF FIT

Parameters such as the mean and the standard deviation are the most common way to describe a population, but there are situations in which a researcher has questions about the proportions or relative frequencies for a distribution. For example,

How does the number of women lawyers compare with the number of men in the profession?

Of the two leading brands of cola, which is preferred by most Americans?

In the past 10 years, has there been a significant change in the proportion of college students who declare a business major?

The name of the test comes from the Greek letter χ (chi, pronounced "kye"), which is used to identify the test statistic.

Note that each of the preceding examples asks a question about proportions in the population. In particular, we are not measuring a numerical score for each individual. Instead, the individuals are simply classified into categories and we want to know what proportion of the population is in each category. The *chi-square test for goodness of fit* is specifically designed to answer this type of question. In general terms, this chi-square test uses the proportions obtained for sample data to test hypotheses about the corresponding proportions in the population.

DEFINITION

The **chi-square test for goodness of fit** uses sample data to test hypotheses about the shape or proportions of a population distribution. The test determines how well the obtained sample proportions fit the population proportions specified by the null hypothesis.

Recall from Chapter 2 that a frequency distribution is defined as a tabulation of the number of individuals located in each category of the scale of measurement. In a frequency distribution graph, the categories that make up the scale of measurement are listed on the X-axis. In a frequency distribution table, the categories are listed in the first column. With chi-square tests, however, it is customary to present the scale of measurement as a series of boxes, with each box corresponding to a separate category

on the scale. The frequency corresponding to each category is simply presented as a number written inside the box. Figure 17.1 shows how a distribution of eye colors for a set of $n = 40$ students can be presented as a graph, a table, or a series of boxes. The scale of measurement for this example consists of four categories of eye color (blue, brown, green, other).

THE NULL HYPOTHESIS FOR THE GOODNESS-OF-FIT TEST

For the chi-square test of goodness of fit, the null hypothesis specifies the proportion (or percentage) of the population in each category. For example, a hypothesis might state that 50% of all lawyers are men and 50% are women. The simplest way of presenting this hypothesis is to put the hypothesized proportions in the series of boxes representing the scale of measurement:

	Men	Women
H_0:	50%	50%

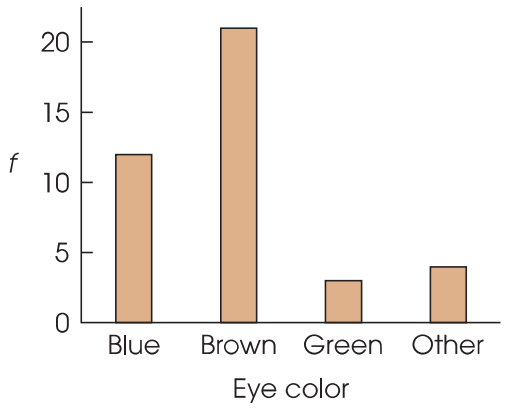
Although it is conceivable that a researcher could choose any proportions for the null hypothesis, there usually is some well-defined rationale for stating a null hypothesis. Generally H_0 falls into one of the following categories:

1. **No Preference, Equal Proportions.** The null hypothesis often states that there is no preference among the different categories. In this case, H_0 states that the population is divided equally among the categories. For example, a hypothesis stating that there is no preference among the three leading brands of soft drinks would specify a population distribution as follows:

	Brand X	Brand Y	Brand Z
H_0:	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

(Preferences in the population are equally divided among the three soft drinks.)

The no-preference hypothesis is used in situations in which a researcher wants to determine whether there are any preferences among the categories, or whether the proportions differ from one category to another.



Eye color (X)	f
Blue	12
Brown	21
Green	3
Other	4

Blue	Brown	Green	Other
12	21	3	4

FIGURE 17.1

Distribution of eye colors for a sample of $n = 40$ individuals. The same frequency distribution is shown as a bar graph, as a table, and with the frequencies written in a series of boxes.

Because the null hypothesis for the goodness-of-fit test specifies an exact distribution for the population, the alternative hypothesis (H_1) simply states that the population distribution has a different shape from that specified in H_0. If the null hypothesis states that the population is equally divided among three categories, then the alternative hypothesis says that the population is not divided equally.

2. **No Difference from a Known Population.** The null hypothesis can state that the proportions for one population are not different from the proportions that are known to exist for another population. For example, suppose it is known that 28% of the licensed drivers in the state are younger than 30 years old and 72% are 30 or older. A researcher might wonder whether this same proportion holds for the distribution of speeding tickets. The null hypothesis would state that tickets are handed out equally across the population of drivers, so there is no difference between the age distribution for drivers and the age distribution for speeding tickets. Specifically, the null hypothesis would be

	Tickets Given to Drivers Younger Than 30	Tickets Given to Drivers 30 or Older
H_0:	28%	72%

(Proportions for the population of tickets are not different from proportions for drivers.)

The no-difference hypothesis is used when a specific population distribution is already known. For example, you may have a known distribution from an earlier time, and the question is whether there has been any change in the proportions. Or, you may have a known distribution for one population (drivers) and the question is whether a second population (speeding tickets) has the same proportions.

Again, the alternative hypothesis (H_1) simply states that the population proportions are not equal to the values specified by the null hypothesis. For this example, H_1 would state that the number of speeding tickets is disproportionately high for one age group and disproportionately low for the other.

THE DATA FOR THE GOODNESS-OF-FIT TEST

The data for a chi-square test are remarkably simple. There is no need to calculate a sample mean or SS; you just select a sample of n individuals and count how many are in each category. The resulting values are called observed frequencies. The symbol for observed frequency is f_o. For example, the following data represent observed frequencies for a sample of 40 college students. The students were classified into three categories based on the number of times they reported exercising each week.

No Exercise	1 Time a Week	More Than Once a Week
15	19	6

$n = 40$

Notice that each individual in the sample is classified into one and only one of the categories. Thus, the frequencies in this example represent three completely separate groups of students: 15 who do not exercise regularly, 19 who average once a week, and 6 who exercise more than once a week. Also note that the observed frequencies add up to the total sample size: $\sum f_o = n$. Finally, you should realize that we are not assigning individuals to categories. Instead, we are simply measuring individuals to determine the category in which they belong.

DEFINITION

The **observed frequency** is the number of individuals from the sample who are classified in a particular category. Each individual is counted in one and only one category.

EXPECTED FREQUENCIES

The general goal of the chi-square test for goodness of fit is to compare the data (the observed frequencies) with the null hypothesis. The problem is to determine how well the data fit the distribution specified in H_0—hence the name *goodness of fit*.

The first step in the chi-square test is to construct a hypothetical sample that represents how the sample distribution would look if it were in perfect agreement with the proportions stated in the null hypothesis. Suppose, for example, the null hypothesis states that the population is distributed in three categories with the following proportions:

	Category A	Category B	Category C
H_0:	25%	50%	25%

(The population is distributed across the three categories with 25% in category A, 50% in category B, and 25% in category C.)

If this hypothesis is correct, how would you expect a random sample of $n = 40$ individuals to be distributed among the three categories? It should be clear that your best strategy is to predict that 25% of the sample would be in category A, 50% would be in category B, and 25% would be in category C. To find the exact frequency expected for each category, multiply the sample size (n) by the proportion (or percentage) from the null hypothesis. For this example, you would expect

$$25\% \text{ of } 40 = 0.25(40) = 10 \text{ individuals in category A}$$

$$50\% \text{ of } 40 = 0.50(40) = 20 \text{ individuals in category B}$$

$$25\% \text{ of } 40 = 0.25(40) = 10 \text{ individuals in category C}$$

The frequency values predicted from the null hypothesis are called *expected frequencies*. The symbol for expected frequency is f_e, and the expected frequency for each category is computed by

$$\text{expected frequency} = f_e = pn \tag{17.1}$$

where p is the proportion stated in the null hypothesis and n is the sample size.

DEFINITION

The **expected frequency** for each category is the frequency value that is predicted from the proportions in the null hypothesis and the sample size (n). The expected frequencies define an ideal, *hypothetical* sample distribution that would be obtained if the sample proportions were in perfect agreement with the proportions specified in the null hypothesis.

Note that the no-preference null hypothesis always produces equal f_e values for all categories because the proportions (p) are the same for all categories. On the other hand, the no-difference null hypothesis typically does not produce equal values for the expected frequencies because the hypothesized proportions typically vary from one category to another. You also should note that the expected frequencies are calculated, hypothetical values and the numbers that you obtain may be decimals or fractions. The observed frequencies, on the other hand, always represent real individuals and always are whole numbers.

THE CHI-SQUARE STATISTIC

The general purpose of any hypothesis test is to determine whether the sample data support or refute a hypothesis about the population. In the chi-square test for goodness of fit, the sample is expressed as a set of observed frequencies (f_o values), and the null hypothesis is used to generate a set of expected frequencies (f_e values). The *chi-square statistic* simply measures how well the data (f_o) fit the hypothesis (f_e). The symbol for the chi-square statistic is χ^2. The formula for the chi-square statistic is

$$\text{chi-square} = \chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e} \tag{17.2}$$

As the formula indicates, the value of chi-square is computed by the following steps:

1. Find the difference between f_o (the data) and f_e (the hypothesis) for each category.
2. Square the difference. This ensures that all values are positive.
3. Next, divide the squared difference by f_e.
4. Finally, add the values from all of the categories.

The first two steps determine the numerator of the chi-square statistic and should be easy to understand. Specifically, the numerator measures how much difference there is between the data (the f_o values) and the hypothesis (represented by the f_e values). The final step is also reasonable: we add the values to obtain the total discrepancy between the data and the hypothesis. Thus, a large value for chi-square indicates that the data do not fit the hypothesis, and leads us to reject the null hypothesis.

However, the third step, which determines the denominator of the chi-square statistic, is not so obvious. Why must we divide by f_e before we add the category values? The answer to this question is that the obtained discrepancy between f_o and f_e is viewed as *relatively* large or *relatively* small depending on the size of the expected frequency. This point is demonstrated in the following analogy.

Suppose that you were going to throw a party and you *expected* 1,000 people to show up. However, at the party you counted the number of guests and *observed* that 1,040 actually showed up. Forty more guests than expected are no major problem when all along you were planning for 1,000. There will still probably be enough beer and potato chips for everyone. On the other hand, suppose you had a party and you expected 10 people to attend but instead 50 actually showed up. Forty more guests in this case spell big trouble. How "significant" the discrepancy is depends in part on what you were originally expecting. With very large expected frequencies, allowances are made for more error between f_o and f_e. This is accomplished in the chi-square formula by dividing the squared discrepancy for each category, $(f_o - f_e)^2$, by its expected frequency.

THE CHI-SQUARE DISTRIBUTION AND DEGREES OF FREEDOM

It should be clear from the chi-square formula that the numerical value of chi-square is a measure of the discrepancy between the observed frequencies (data) and the expected frequencies (H_0). As usual, the sample data are not expected to provide a perfectly accurate representation of the population. In this case, the proportions or observed frequencies in the sample are not expected to be exactly equal to the proportions in the population. Thus, if there are small discrepancies between the f_o and f_e values, we obtain a small value for chi-square and we conclude that there is a good fit between the data and the hypothesis (fail to reject H_0). However, when there

are large discrepancies between f_o and f_e, we obtain a large value for chi-square and conclude that the data do not fit the hypothesis (reject H_0). To decide whether a particular chi-square value is "large" or "small," we must refer to a *chi-square distribution*. This distribution is the set of chi-square values for all of the possible random samples when H_0 is true. Much like other distributions that we have examined (t distribution, F distribution), the chi-square distribution is a theoretical distribution with well-defined characteristics. Some of these characteristics are easy to infer from the chi-square formula.

1. The formula for chi-square involves adding squared values, so you can never obtain a negative value. Thus, all chi-square values are zero or larger.
2. When H_0 is true, you expect the data (f_o values) to be close to the hypothesis (f_e values). Thus, we expect chi-square values to be small when H_0 is true.

These two factors suggest that the typical chi-square distribution is positively skewed (Figure 17.2). Note that small values, near zero, are expected when H_0 is true and large values (in the right-hand tail) are very unlikely. Thus, unusually large values of chi-square form the critical region for the hypothesis test.

Although the typical chi-square distribution is positively skewed, there is one other factor that plays a role in the exact shape of the chi-square distribution—the number of categories. Recall that the chi-square formula requires that you add values from every category. The more categories you have, the more likely it is that you will obtain a large sum for the chi-square value. On average, chi-square is larger when you are adding values from 10 categories than when you are adding values from only 3 categories. As a result, there is a whole family of chi-square distributions, with the exact shape of each distribution determined by the number of categories used in the study. Technically, each specific chi-square distribution is identified by degrees of freedom (df) rather than the number of categories. For the goodness-of-fit test, the degrees of freedom are determined by

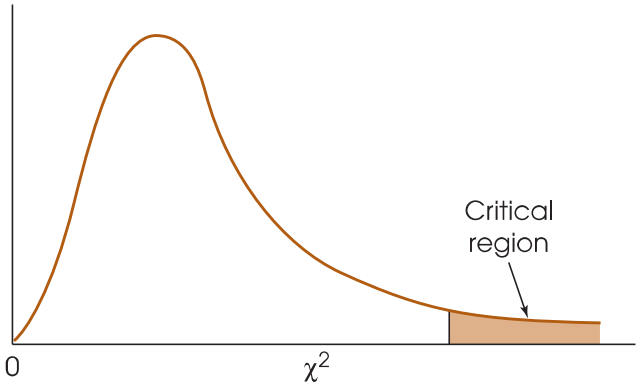
$$df = C - 1 \tag{17.3}$$

Caution: The *df* for a chi-square test is *not* related to sample size (n), as it is in most other tests.

where C is the number of categories. A brief discussion of this df formula is presented in Box 17.1. Figure 17.3 shows the general relationship between df and the shape of the chi-square distribution. Note that the chi-square values tend to get larger (shift to the right) as the number of categories and the degrees of freedom increase.

FIGURE 17.2

Chi-square distributions are positively skewed. The critical region is placed in the extreme tail, which reflects large chi-square values.



BOX 17.1 A CLOSER LOOK AT DEGREES OF FREEDOM

Degrees of freedom for the chi-square test literally measure the number of free choices that exist when you are determining the null hypothesis or the expected frequencies. For example, when you are classifying individuals into three categories, you have exactly two free choices in stating the null hypothesis. You may select any two proportions for the first two categories, but then the third proportion is determined. If you hypothesize 25% in the first category and 50% in the second category, then the third category must be 25% to account for 100% of the population.

Category A	Category B	Category C
10	20	?

In general, you are free to select proportions for all but one of the categories, but then the final proportion is determined by the fact that the entire set must total 100%. Thus, you have $C - 1$ free choices, where C is the number of categories: degrees of freedom, *df*, equal $C - 1$.

LOCATING THE CRITICAL REGION FOR A CHI-SQUARE TEST

Recall that a large value for the chi-square statistic indicates a big discrepancy between the data and the hypothesis, and suggests that we reject H_0. To determine whether a particular chi-square value is significantly large, you must consult the table entitled The Chi-Square Distribution (Appendix B). A portion of the chi-square table is shown in Table 17.2. The first column lists *df* values for the chi-square test, and the other column heads are proportions (alpha levels) in the extreme right-hand tail of the distribution. The numbers in the body of the table are the critical values of chi-square. The table shows, for example, that when the null hypothesis is true and $df = 3$, only 5% (.05) of the chi-square values are greater than 7.81, and only 1% (.01) are greater than 11.34. Thus, with $df = 3$, any chi-square value greater than 7.81 has a probability of $p < .05$, and any value greater than 11.34 has a probability of $p < .01$.

FIGURE 17.3

The shape of the chi-square distribution for different values of *df*. As the number of categories increases, the peak (mode) of the distribution has a larger chi-square value.

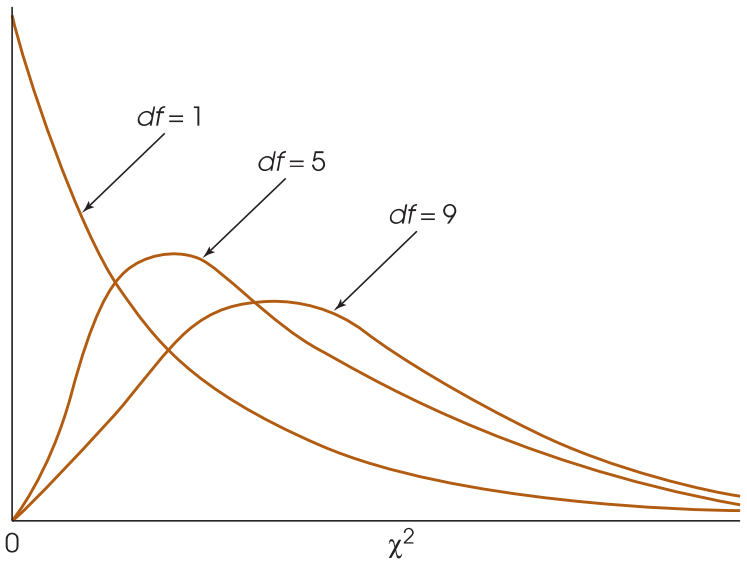


TABLE 17.2

A portion of the table of critical values for the chi-square distribution.

	Proportion in Critical Region				
df	0.10	0.05	0.025	0.01	0.005
1	2.71	3.84	5.02	6.63	7.88
2	4.61	5.99	7.38	9.21	10.60
3	6.25	7.81	9.35	11.34	12.84
4	7.78	9.49	11.14	13.28	14.86
5	9.24	11.07	12.83	15.09	16.75
6	10.64	12.59	14.45	16.81	18.55
7	12.02	14.07	16.01	18.48	20.28
8	13.36	15.51	17.53	20.09	21.96
9	14.68	16.92	19.02	21.67	23.59

EXAMPLE OF THE CHI-SQUARE TEST FOR GOODNESS OF FIT

We use the same step-by-step process for testing hypotheses with chi-square as we used for other hypothesis tests. In general, the steps consist of stating the hypotheses, locating the critical region, computing the test statistic, and making a decision about H_0. The following example demonstrates the complete process of hypothesis testing with the goodness-of-fit test.

EXAMPLE 17.1

A psychologist examining art appreciation selected an abstract painting that had no obvious top or bottom. Hangers were placed on the painting so that it could be hung with any one of the four sides at the top. The painting was shown to a sample of $n = 50$ participants, and each was asked to hang the painting in the orientation that looked correct. The following data indicate how many people chose each of the four sides to be placed at the top:

Top Up (Correct)	Bottom Up	Left Side Up	Right Side Up
18	17	7	8

The question for the hypothesis test is whether there are any preferences among the four possible orientations. Are any of the orientations selected more (or less) often than would be expected simply by chance?

STEP 1

State the hypotheses and select an alpha level. The hypotheses can be stated as follows:

H_0: In the general population, there is no preference for any specific orientation. Thus, the four possible orientations are selected equally often, and the population distribution has the following proportions:

Top Up (Correct)	Bottom Up	Left Side Up	Right Side Up
25%	25%	25%	25%

H_1: In the general population, one or more of the orientations is preferred over the others.

We use $\alpha = .05$.

STEP 2 Locate the critical region. For this example, the value for degrees of freedom is

$$df = C - 1 = 4 - 1 = 3$$

For $df = 3$ and $\alpha = .05$, the table of critical values for chi-square indicates that the critical χ^2 has a value of 7.81. The critical region is sketched in Figure 17.4.

STEP 3 Calculate the chi-square statistic. The calculation of chi-square is actually a two-stage process. First, you must compute the expected frequencies from H_0 and then calculate the value of the chi-square statistic. For this example, the null hypothesis specifies that one-quarter of the population ($p = 25\%$) will be in each of the four categories. According to this hypothesis, we should expect one-quarter of the sample to be in each category. With a sample of $n = 50$ individuals, the expected frequency for each category is

Expected frequencies are computed and may be decimal values. Observed frequencies are always whole numbers.

$$f_e = pn = \frac{1}{4}(50) = 12.5$$

The observed frequencies and the expected frequencies are presented in Table 17.3. Using these values, the chi-square statistic may now be calculated.

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$
$$= \frac{(18-12.5)^2}{12.5} + \frac{(17-12.5)^2}{12.5} + \frac{(7-12.5)^2}{12.5} + \frac{(8-12.5)^2}{12.5}$$
$$= \frac{30.25}{12.5} + \frac{20.25}{12.5} + \frac{30.25}{12.5} + \frac{20.25}{12.5}$$
$$= 2.42 + 1.62 + 2.42 + 1.62$$
$$= 8.08$$

STEP 4 State a decision and a conclusion. The obtained chi-square value is in the critical region. Therefore, H_0 is rejected, and the researcher may conclude that the four orientations are not equally likely to be preferred. Instead, there are significant differences among the four orientations, with some selected more often and others less often than would be expected by chance.

FIGURE 17.4

For Example 17.1, the critical region begins at a chi-square value of 7.81.

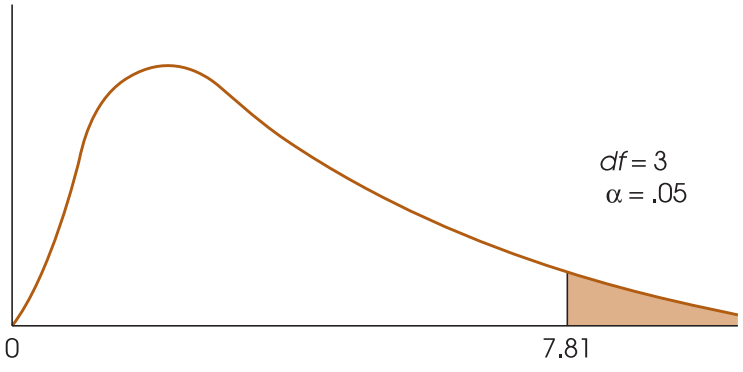


TABLE 17.3

The observed frequencies and the expected frequencies for the chi-square test in Example 17.1.

	Top Up (Correct)	Bottom Up	Left Side Up	Right Side Up
Observed Frequencies	18	17	7	8
Expected Frequencies	12.5	12.5	12.5	12.5

IN THE LITERATURE

REPORTING THE RESULTS FOR CHI-SQUARE

APA style specifies the format for reporting the chi-square statistic in scientific journals. For the results of Example 17.1, the report might state:

> The participants showed significant preferences among the four orientations for hanging the painting, $\chi^2(3, n = 50) = 8.08, p < .05$.

Note that the form of the report is similar to that of other statistical tests we have examined. Degrees of freedom are indicated in parentheses following the chi-square symbol. Also contained in the parentheses is the sample size (n). This additional information is important because the degrees of freedom value is based on the number of categories (C), not sample size. Next, the calculated value of chi-square is presented, followed by the probability that a Type I error has been committed. Because we obtained an extreme, very unlikely value for the chi-square statistic, the probability is reported as *less than* the alpha level. Additionally, the report may provide the observed frequencies (f_o) for each category. This information may be presented in a simple sentence or in a table.

GOODNESS OF FIT AND THE SINGLE-SAMPLE *t* TEST

We began this chapter with a general discussion of the difference between parametric tests and nonparametric tests. In this context, the chi-square test for goodness of fit is an example of a nonparametric test; that is, it makes no assumptions about the parameters of the population distribution, and it does not require data from an interval or ratio scale. In contrast, the single-sample t test introduced in Chapter 9 is an example of a parametric test: It assumes a normal population, it tests hypotheses about the population mean (a parameter), and it requires numerical scores that can be added, squared, divided, and so on.

Although the chi-square test and the single-sample t are clearly distinct, they are also very similar. In particular, both tests are intended to use the data from a single sample to test hypotheses about a single population.

The primary factor that determines whether you should use the chi-square test or the t test is the type of measurement that is obtained for each participant. If the sample data consist of numerical scores (from an interval or ratio scale), it is appropriate to compute a sample mean and use a t test to evaluate a hypothesis about the population mean. For example, a researcher could measure the IQ for each individual in a sample of registered voters. A t test could then be used to evaluate a hypothesis about the mean IQ for the entire population of registered voters. On the other hand, if the individuals in the sample are classified into nonnumerical categories (on a nominal or ordinal scale), then the researcher would use a chi-square test to evaluate a hypothesis about the population proportions. For example, a researcher could classify people according to gender by simply counting the number of males and females in a sample of registered voters. A chi-square test would then be appropriate to evaluate a hypothesis about the population proportions.

LEARNING CHECK

1. For a chi-square test, the observed frequencies are always whole numbers. (True or false?)
2. For a chi-square test, the expected frequencies are always whole numbers. (True or false?)
3. A researcher has developed three different designs for a computer keyboard. A sample of $n = 60$ participants is obtained, and each individual tests all three keyboards and identifies his or her favorite. The frequency distribution of preferences is as follows:

Design A	Design B	Design C
23	12	25

$n = 60$

   a. What is the *df* value for the chi-square statistic?
   b. Assuming that the null hypothesis states that there are no preferences among the three designs, find the expected frequencies for the chi-square test.

ANSWERS

1. True. Observed frequencies are obtained by counting people in the sample.
2. False. Expected frequencies are computed and may be fractions or decimal values.
3. a. $df = 2$
   b. According to the null hypothesis one-third of the population would prefer each design. The expected frequencies should show one-third of the sample preferring each design. The expected frequencies are all 20.

17.3 THE CHI-SQUARE TEST FOR INDEPENDENCE

The chi-square statistic may also be used to test whether there is a relationship between two variables. In this situation, each individual in the sample is measured or classified on two separate variables. For example, a group of students could be classified in terms of personality (introvert, extrovert) and in terms of color preference (red, yellow, green, or blue). Usually, the data from this classification are presented in the form of a matrix, where the rows correspond to the categories of one variable and the columns correspond to the categories of the second variable. Table 17.4 presents hypothetical data for a sample of $n = 200$ students who have been classified by personality and color preference. The number in each box, or cell, of the matrix indicates the frequency, or number of individuals in that particular group. In Table 17.4, for example, there are 10 students who were classified as introverted and who selected red as their preferred color. To obtain these data, the researcher first selects a random sample of $n = 200$ students. Each student is then given a personality test and is asked to select a preferred color from among the four choices. Note that the classification is based on the measurements for each student; the researcher does not assign students to categories. Also, note that the data consist of frequencies, not scores, from a sample. The goal is to use the frequencies from the sample to test a hypothesis about the population frequency distribution. Specifically, are these data sufficient to conclude that there is a significant relationship between personality and color preference in the population of students?

You should realize that the color preference study shown in Table 17.3 is an example of nonexperimental research (Chapter 1, page 17). The researcher did not manipulate any variable and the participants were not randomly assigned to groups or

TABLE 17.4
Color preferences according to personality types.

	Red	Yellow	Green	Blue	
Introvert	10	3	15	22	50
Extrovert	90	17	25	18	150
	100	20	40	40	$n = 200$

treatment conditions. However, similar data are often obtained from true experiments. A good example is the study described in the Preview, in which Loftus and Palmer (1974) demonstrate how eyewitness memory can be influenced by the kinds of questions that witnesses are asked. In the study, a sample of 150 students watched a film of an automobile accident. After watching the film, the students were separated into three groups and questioned about the accident. The researchers manipulated the type of question each group was asked. One group was asked to estimate the speed of the cars when they "smashed into each other." Another group estimated speed when the cars "hit each other." A third group served as a control and was not asked any question about the speed of the two cars. A week later, the participants returned and were asked if they remembered seeing any broken glass in the accident. (There was no broken glass in the film.) The researchers recorded the number of Yes and No responses for each group (see Table 17.1, page 592). As with the color preference data, the researchers would like to use the frequencies from the sample to test a hypothesis about the corresponding frequency distribution in the population. In this case, the researchers would like to know whether the sample data provide enough evidence to conclude that there is a significant relationship between eyewitnesses' memories and the questions they were asked.

The procedure for using sample frequencies to evaluate hypotheses concerning relationships between variables involves another test using the chi-square statistic. In this situation, however, the test is called the *chi-square test for independence.*

DEFINITION

The **chi-square test for independence** uses the frequency data from a sample to evaluate the relationship between two variables in the population. Each individual in the sample is classified on both of the two variables, creating a two-dimensional frequency-distribution matrix. The frequency distribution for the sample is then used to test hypotheses about the corresponding frequency distribution for the population.

THE NULL HYPOTHESIS FOR THE CHI-SQUARE TEST FOR INDEPENDENCE

The null hypothesis for the chi-square test for independence states that the two variables being measured are independent; that is, for each individual, the value obtained for one variable is not related to (or influenced by) the value for the second variable. This general hypothesis can be expressed in two different conceptual forms, each viewing the data and the test from slightly different perspectives. The data in Table 17.4 describing color preference and personality are used to present both versions of the null hypothesis.

H_0 version 1 For this version of H_0, the data are viewed as a single sample with each individual measured on two variables. The goal of the chi-square test is to evaluate the relationship between the two variables. For the example we are considering, the goal is to determine whether there is a consistent, predictable relationship between personality and color preference. That is, if I know your personality, will it help me to predict your color preference? The null hypothesis states that there is no relationship. The alternative hypothesis, H_1, states that there is a relationship between the two variables.

H_0: For the general population of students, there is no relationship between color preference and personality.

This version of H_0 demonstrates the similarity between the chi-square test for independence and a correlation. In each case, the data consist of two measurements (X and Y) for each individual, and the goal is to evaluate the relationship between the two variables. The correlation, however, requires numerical scores for X and Y. The chi-square test, on the other hand, simply uses frequencies for individuals classified into categories.

H_0 version 2 For this version of H_0, the data are viewed as two (or more) separate samples representing two (or more) populations or treatment conditions. The goal of the chi-square test is to determine whether there are significant differences between the populations. For the example we are considering, the data in Table 17.4 would be viewed as a sample of $n = 50$ introverts (top row) and a separate sample of $n = 150$ extroverts (bottom row). The chi-square test determines whether the distribution of color preferences for introverts is significantly different from the distribution of color preferences for extroverts. From this perspective, the null hypothesis is stated as follows:

H_0: In the population of students, the proportions in the distribution of color preferences for introverts are not different from the proportions in the distribution of color preferences for extroverts. The two distributions have the same shape (same proportions).

This version of H_0 demonstrates the similarity between the chi-square test and an independent-measures t test (or ANOVA). In each case, the data consist of two (or more) separate samples that are being used to test for differences between two (or more) populations. The t test (or ANOVA) requires numerical scores to compute means and mean differences. However, the chi-square test simply uses frequencies for individuals classified into categories. The null hypothesis for the chi-square test states that the populations have the same proportions (same shape). The alternative hypothesis, H_1, simply states that the populations have different proportions. For the example we are considering, H_1 states that the shape of the distribution of color preferences for introverts is different from the shape of the distribution of color preferences for extroverts.

Equivalence of H_0 version 1 and H_0 version 2 Although we have presented two different statements of the null hypothesis, these two versions are equivalent. The first version of H_0 states that color preference is not related to personality. If this hypothesis is correct, then the distribution of color preferences should not depend on personality. In other words, the distribution of color preferences should have the same proportions for introverts and for extroverts, which is the second version of H_0.

For example, if we found that 60% of the introverts preferred red, then H_0 would predict that we also should find that 60% of the extroverts prefer red. In this case, knowing that an individual prefers red does not help you predict his or her personality. Note that finding the *same proportions* indicates *no relationship*.

On the other hand, if the proportions were different, it would suggest that there is a relationship. For example, if red is preferred by 60% of the extroverts but only 10% of the introverts, then there is a clear, predictable relationship between personality and color preference. (If I know your personality, then I can predict your color preference.) Thus, finding *different proportions* means that there is *a relationship* between the two variables.

DEFINITION

Two variables are **independent** when there is no consistent, predictable relationship between them. In this case, the frequency distribution for one variable is not related to (or dependent on) the categories of the second variable. As a result, when two variables are independent, the frequency distribution for one variable has the same shape (same proportions) for all categories of the second variable.

Thus, stating that there is no relationship between two variables (version 1 of H_0) is equivalent to stating that the distributions have equal proportions (version 2 of H_0).

OBSERVED AND EXPECTED FREQUENCIES

The chi-square test for independence uses the same basic logic that was used for the goodness-of-fit test. First, a sample is selected and each individual is classified or categorized. Because the test for independence considers two variables, every individual is classified on both variables, and the resulting frequency distribution is presented as a two-dimensional matrix (see Table 17.4). As before, the frequencies in the sample distribution are called *observed frequencies* and are identified by the symbol f_o.

The next step is to find the expected frequencies, or f_e values, for this chi-square test. As before, the e*xpected frequencies* define an ideal hypothetical distribution that is in perfect agreement with the null hypothesis. Once the expected frequencies are obtained, we compute a chi-square statistic to determine how well the data (observed frequencies) fit the null hypothesis (expected frequencies).

Although you can use either version of the null hypothesis to find the expected frequencies, the logic of the process is much easier when you use H_0 stated in terms of equal proportions. For the example we are considering, the null hypothesis states

H_0: The frequency distribution of color preference has the same shape (same proportions) for both categories of personality.

To find the expected frequencies, we first determine the overall distribution of color preferences and then apply this distribution to both categories of personality. Table 17.5 shows an empty matrix corresponding to the data from Table 17.4. Notice that the empty matrix includes all of the row totals and column totals from the original sample data. The row totals and column totals are essential for computing the expected frequencies.

The column totals for the matrix describe the overall distribution of color preferences. For these data, 100 people selected red as their preferred color. Because the total sample consists of 200 people, the proportion selecting red is 100 out of 200, or 50%. The complete set of color preference proportions is as follows:

100 out of 200 = 50% prefer red

20 out of 200 = 10% prefer yellow

40 out of 200 = 20% prefer green

40 out of 200 = 20% prefer blue

The row totals in the matrix define the two samples of personality types. For example, the matrix in Table 17.5 shows a total of 50 introverts (the top row) and a sample of 150 extroverts (the bottom row). According to the null hypothesis, both personality groups should have the same proportions for color preferences. To find the expected frequencies, we simply apply the overall distribution of color preferences to

TABLE 17.5

An empty frequency distribution matrix showing only the row totals and column totals. (These numbers describe the basic characteristics of the sample from Table 17.4.)

	Red	Yellow	Green	Blue	
Introvert					50
Extrovert					150
	100	20	40	40	

each sample. Beginning with the sample of 50 introverts in the top row, we obtain expected frequencies of

50% prefer red: f_e = 50% of 50 = 0.50(50) = 25
10% prefer yellow: f_e = 10% of 50 = 0.10(50) = 5
20% prefer green: f_e = 20% of 50 = 0.20(50) = 10
20% prefer blue: f_e = 20% of 50 = 0.20(50) = 10

Using exactly the same proportions for the sample of $n = 150$ extroverts in the bottom row, we obtain expected frequencies of

50% prefer red: f_e = 50% of 150 = 0.50(50) = 75
10% prefer yellow: f_e = 10% of 150 = 0.10(50) = 15
20% prefer green: f_e = 20% of 150 = 0.20(50) = 30
20% prefer blue: f_e = 20% of 150 = 0.20(50) = 30

The complete set of expected frequencies is shown in Table 17.6. Notice that the row totals and the column totals for the expected frequencies are the same as those for the original data (the observed frequencies) in Table 17.3.

A simple formula for determining expected frequencies Although expected frequencies are derived directly from the null hypothesis and the sample characteristics, it is not necessary to go through extensive calculations to find f_e values. In fact, there is a simple formula that determines f_e for any cell in the frequency distribution matrix:

$$f_e = \frac{f_c f_r}{n} \tag{17.4}$$

where f_c is the frequency total for the column (column total), f_r is the frequency total for the row (row total), and n is the number of individuals in the entire sample. To demonstrate this formula, we compute the expected frequency for introverts selecting yellow in Table 17.6. First, note that this cell is located in the top row and second column in the table. The column total is $f_c = 20$, the row total is $f_r = 50$, and the sample size is $n = 200$. Using these values in formula 17.4, we obtain

$$f_e = \frac{f_c f_r}{n} = \frac{20(50)}{200} = 5$$

TABLE 17.6

Expected frequencies corresponding to the data in Table 17.4. (This is the distribution predicted by the null hypothesis.)

	Red	Yellow	Green	Blue	
Introvert	25	5	10	10	50
Extrovert	75	15	30	30	150
	100	20	40	40	

This is identical to the expected frequency we obtained using percentages from the overall distribution.

THE CHI-SQUARE STATISTIC AND DEGREES OF FREEDOM

The chi-square test for independence uses exactly the same chi-square formula as the test for goodness of fit:

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

As before, the formula measures the discrepancy between the data (f_o values) and the hypothesis (f_e values). A large discrepancy produces a large value for chi-square and indicates that H_0 should be rejected. To determine whether a particular chi-square statistic is significantly large, you must first determine degrees of freedom (*df*) for the statistic and then consult the chi-square distribution in Appendix B. For the chi-square test of independence, degrees of freedom are based on the number of cells for which you can freely choose expected frequencies. Recall that the f_e values are partially determined by the sample size (n) and by the row totals and column totals from the original data. These various totals restrict your freedom in selecting expected frequencies. This point is illustrated in Table 17.7. Once three of the f_e values have been selected, all of the other f_e values in the table are also determined. For example, the bottom number in the first column must be 75 to produce a column total of 100. Similarly, the last number in the top row must be 10 to produce a row total of 50. In general, the row totals and the column totals restrict the final choices in each row and column. As a result, we may freely choose all but one f_e in each row and all but one f_e in each column. If R is the number of rows and C is the number of columns, and you remove the last column and the bottom row from the matrix, you are left with a smaller matrix that has $C - 1$ columns and $R - 1$ rows. The number of cells in the smaller matrix determines the *df* value. Thus, the total number of f_e values that you can freely choose is $(R - 1)(C - 1)$, and the degrees of freedom for the chi-square test of independence are given by the formula

$$df = (R - 1)(C - 1) \tag{17.5}$$

Also note that once you calculate the expected frequencies to fill the smaller matrix, the rest of the f_e values can be found by subtraction.

AN EXAMPLE OF THE CHI-SQUARE TEST FOR INDEPENDENCE

The following example demonstrates the complete hypothesis-testing procedure for the chi-square test for independence.

TABLE 17.7

Degrees of freedom and expected frequencies. (Once three values have been selected, all the remaining expected frequencies are determined by the row totals and the column totals. This example has only three free choices, so $df = 3$.)

Red	Yellow	Green	Blue	
25	5	10	?	50
?	?	?	?	150
100	20	40	40	

EXAMPLE 17.2 Research has demonstrated strong gender differences in teenagers' approaches to dealing with mental health issues (Chandra & Minkovitz, 2006). In a typical study, eighth-grade students are asked to report their willingness to use mental health services in the event they were experiencing emotional or other mental health problems. Typical data for a sample of $n = 150$ students are shown in Table 17.8. Do the data show a significant relationship between gender and willingness to seek mental health assistance?

STEP 1 State the hypotheses, and select a level of significance. According to the null hypothesis, the two variables are independent. This general hypothesis can be stated in two different ways:

Version 1

H_0: In the general population, there is no relationship between gender and willingness to use mental health services.

This version of H_0 emphasizes the similarity between the chi-square test and a correlation. The corresponding alternative hypothesis would state:

H_1: In the general population, there is a consistent, predictable relationship between gender and willingness to use mental health services.

Version 2

H_0: In the general population, the distribution of reported willingness to use mental health services has the same proportions for males and for females.

The corresponding alternative hypothesis would state:

H_1: In the general population, the distribution of reported willingness to use mental health services for males has proportions that are different from those in the distribution for females.

The second version of H_0 emphasizes the similarity between the chi-square test and the independent-measures t test.

Remember that the two versions for the hypotheses are equivalent. The choice between them is largely determined by how the researcher wants to describe the outcome. For example, a researcher may want to emphasize the *relationship* between variables or the *difference* between groups.

For this test, we use $\alpha = .05$.

STEP 2 Determine the degrees of freedom and locate the critical region. For the chi-square test for independence,

$$df = (R - 1)(C - 1) = (2 - 1)(3 - 1) = 2$$

With $df = 2$ and $\alpha = .05$, the critical value for chi-square is 5.99 (see Table B.8, p. 711).

TABLE 17.8

A frequency distribution showing willingness to use mental health services according to gender for a sample of $n = 150$ students.

	Willingness to Use Mental Health Services			
	Probably No	Maybe	Probably Yes	
Males	17	32	11	60
Females	13	43	34	90
	30	75	45	$n = 150$

STEP 3 Determine the expected frequencies, and compute the chi-square statistic. The following table shows an empty matrix with the same row totals and column totals as the original data. The expected frequencies must maintain the same row totals and column totals, and create an ideal frequency distribution that perfectly represents the null hypothesis. Specifically, the proportions for the group of 60 males must be the same as the proportions for the group of 90 females.

Willingness to Use Mental Health Services

	Probably No	Maybe	Probably Yes	
Males				60
Females				90
	30	75	45	$n = 150$

The column totals describe the overall distribution of willingness. These totals indicate that 30 out of 150 students reported that they would probably not use mental health services. This proportion corresponds to $\frac{30}{150}$, or 20% of the total sample. Similarly, $\frac{75}{150} = 50\%$ reported that they might use mental health services. Finally, $\frac{45}{150} = 30\%$ reported that they probably would use the services. The null hypothesis (version 2) states that these proportions are the same for males and females. Therefore, we simply apply the proportions to each group to obtain the expected frequencies. For the group of 60 males (top row), we obtain

20% of 60 = 12 males who would probably not seek services

50% of 60 = 30 males who might seek services

30% of 60 = 18 males who probably would seek services

For the group of 90 females (bottom row), we expect

20% of 90 = 18 females who would probably not seek services

50% of 90 = 45 females who may seek services

30% of 90 = 27 females who probably would seek services

These expected frequencies are summarized in Table 17.9.

The chi-square statistic is now used to measure the discrepancy between the data (the observed frequencies in Table 17.8) and the null hypothesis that was used to generate the expected frequencies in Table 17.9.

TABLE 17.9

The expected frequencies (f_e values) of willingness to use mental services is completely independent of gender.

Willingness to Use Mental Health Services

	Probably No	Maybe	Probably Yes	
Males	12	30	18	60
Females	18	45	27	90
	30	75	45	

$$\chi^2 = \frac{(17-12)^2}{12} + \frac{(32-30)^2}{30} - \frac{(11-18)^2}{18} + \frac{(13-18)^2}{18} + \frac{(43-45)^2}{45} + \frac{(34-27)^2}{27}$$

$$= 2.08 + 0.13 + 2.72 + 1.39 + 0.09 + 1.82$$

$$= 8.23$$

STEP 4 Make a decision regarding the null hypothesis and the outcome of the study. The obtained chi-square value exceeds the critical value (5.99). Therefore, the decision is to reject the null hypothesis. In the literature, this would be reported as a significant result with $\chi^2(2, n = 150) = 8.23$, $p < .05$. According to version 1 of H_0, this means that we have decided there is a significant relationship between gender and willingness to use mental health services. Expressed in terms of version 2 of H_0, the data show a significant difference between males' and females' attitudes toward using mental health services. To describe the details of the significant result, you must compare the original data (Table 17.8) with the expected frequencies in Table 17.9. Looking at the two tables, it should be clear that males were less willing to use mental health services and females were more willing than would be expected if the two variables were independent.

LEARNING CHECK

1. A researcher would like to know which factors are most important to people who are buying a new car. A sample of $n = 200$ customers between the ages of 20 and 29 are asked to identify the most important factor in the decision process: Performance, Reliability, or Style. The researcher would like to know whether there is a difference between the factors identified by women compared to those identified by men. The data are as follows:

Observed Frequencies of Most Important factor According to Gender

	Performance	Reliability	Style	Totals
Male	21	33	26	80
Female	19	67	34	120
Totals	40	100	60	

a. State the null hypotheses.

b. Determine the value for *df* for the chi-square test.

c. Compute the expected frequencies.

ANSWERS

1. a. H_0: In the population, the distribution of preferred factors for men has the same proportions as the distribution for women.

b. $df = 2$

c. f_e values are as follows:

Expected Frequencies

	Performance	Reliability	Style
Male	16	40	24
Female	24	60	36

17.4 MEASURING EFFECT SIZE FOR THE CHI-SQUARE TEST FOR INDEPENDENCE

A hypothesis test, like the chi-square test for independence, evaluates the statistical significance of the results from a research study. Specifically, the intent of the test is to determine whether it is likely that the patterns or relationships observed in the sample data could have occurred without any corresponding patterns or relationships in the population. Tests of significance are influenced not only by the size or strength of the treatment effects but also by the size of the samples. As a result, even a small effect can be statistically significant if it is observed in a very large sample. Because a significant effect does not necessarily mean a large effect, it is generally recommended that the outcome of a hypothesis test be accompanied by a measure of the effect size. This general recommendation also applies to the chi-square test for independence.

THE PHI-COEFFICIENT AND CRAMÉR'S *V*

In Chapter 15 (p. 545), we introduced the *phi-coefficient* as a measure of correlation for data consisting of two dichotomous variables (both variables have exactly two values). This same situation exists when the data for a chi-square test for independence form a 2×2 matrix (again, each variable has exactly two values). In this case, it is possible to compute the correlation phi (ϕ) in addition to the chi-square hypothesis test for the same set of data. Because phi is a correlation, it measures the strength of the relationship, rather than the significance, and thus provides a measure of effect size. The value for the phi-coefficient can be computed directly from chi-square by the following formula:

Caution: The value of χ^2 is already a squared value. Do not square it again.

$$\phi = \sqrt{\frac{\chi^2}{n}} \tag{17.6}$$

The value of the phi-coefficient is determined entirely by the *proportions* in the 2×2 data matrix and is completely independent of the absolute size of the frequencies. The chi-square value, however, is influenced by the proportions and by the size of the frequencies. This distinction is demonstrated in the following example.

EXAMPLE 17.3

The following data show a frequency distribution evaluating the relationship between gender and preference between two candidates for student president.

Candidate

	A	B
Male	5	10
Female	10	5

Note that the data show that males prefer candidate B by a 2-to-1 margin and females prefer candidate A by 2 to 1. Also note that the sample includes a total of 15 males and 15 females. We will not perform all the arithmetic here, but these data produce chi-square equal to 3.33 (which is not significant) and a phi-coefficient of 0.333.

Next we keep exactly the same proportions in the data, but double all of the frequencies. The resulting data are as follows:

	Candidate A	Candidate B
Male	10	20
Female	20	10

Once again, males prefer candidate B by 2 to 1 and females prefer candidate A by 2 to 1. However, the sample now contains 30 males and 30 females. For these new data, the value of chi-square is 6.66, twice as big as it was before (and now significant with $\alpha = .05$), but the value of the phi-coefficient is still 0.333.

Because the proportions are the same for the two samples, the value of the phi-coefficient is unchanged. However, the larger sample provides more convincing evidence than the smaller sample, so the larger sample is more likely to produce a significant result.

The interpretation of ϕ follows the same standards used to evaluate a correlation (Table 9.3, p. 299 shows the standards for squared correlations): a correlation of 0.10 is a small effect, 0.30 is a medium effect, and 0.50 is a large effect. Occasionally, the value of ϕ is squared (ϕ^2) and is reported as a percentage of variance accounted for, exactly the same as r^2.

When the chi-square test involves a matrix larger than 2×2, a modification of the phi-coefficient, known as *Cramér's* V, can be used to measure effect size.

$$V = \sqrt{\frac{\chi^2}{n(df^*)}} \tag{17.7}$$

Note that the formula for Cramér's V (17.7) is identical to the formula for the phi-coefficient (17.6) except for the addition of df^* in the denominator. The df^* value is *not* the same as the degrees of freedom for the chi-square test, but it is related. Recall that the chi-square test for independence has $df = (R - 1)(C - 1)$, where R is the number of rows in the table and C is the number of columns. For Cramér's V, the value of df^* is the smaller of either $(R - 1)$ or $(C - 1)$.

Cohen (1988) has also suggested standards for interpreting Cramér's V that are shown in Table 17.10. Note that when $df^* = 1$, as in a 2×2 matrix, the criteria for interpreting V are exactly the same as the criteria for interpreting a regular correlation or a phi-coefficient.

We will use the results from Example 17.2 (p. 610) to demonstrate the calculation of Cramér's V. The example evaluated the relationship between gender and willingness to use mental health services. There were two levels of gender and three levels of willingness producing a 2×3 table with a total of $n = 150$ participants. The data produced $\chi^2 = 8.23$. Using these values, we obtain

$$V = \sqrt{\frac{\chi^2}{n(df^*)}} = \sqrt{\frac{8.23}{150(1)}} = \sqrt{0.055} = 0.23$$

TABLE 17.10

Standards for interpreting Cramér's V as proposed by Cohen (1988).

	Small Effect	Medium Effect	Large Effect
For $df^* = 1$	0.10	0.30	0.50
For $df^* = 2$	0.07	0.21	0.35
For $df^* = 3$	0.06	0.17	0.29

According to Cohen's guidelines (see Table 17.10), this value indicates a small or medium relationship.

In a research report, the measure of effect size appears immediately after the results of the hypothesis test. For the study in Example 17.2, the results would be reported as follows:

> The results showed a significant difference between males' and females' attitudes toward using mental health services, $\chi^2(2, n = 50) = 8.23$, $p < .05$, $V = 0.23$.

17.5 ASSUMPTIONS AND RESTRICTIONS FOR CHI-SQUARE TESTS

To use a chi-square test for goodness of fit or a test of independence, several conditions must be satisfied. For any statistical test, violation of assumptions and restrictions casts doubt on the results. For example, the probability of committing a Type I error may be distorted when assumptions of statistical tests are not satisfied. Some important assumptions and restrictions for using chi-square tests are the following:

1. **Independence of Observations.** This is *not* to be confused with the concept of independence between *variables,* as seen in the chi-square test for independence (Section 17.3). One consequence of independent observations is that each observed frequency is generated by a different individual. A chi-square test would be inappropriate if a person could produce responses that can be classified in more than one category or contribute more than one frequency count to a single category. (See p. 254 for more information on independence.)
2. **Size of Expected Frequencies.** A chi-square test should not be performed when the expected frequency of any cell is less than 5. The chi-square statistic can be distorted when f_e is very small. Consider the chi-square computations for a single cell. Suppose that the cell has values of $f_e = 1$ and $f_o = 5$. Note that there is a 4-point difference between the observed and expected frequencies. However, the total contribution of this cell to the total chi-square value is

$$\text{cell} = \frac{(f_o - f_e)^2}{f_e} = \frac{(5 - 1)^2}{1} = \frac{4^2}{1} = 16$$

Now consider another instance, in which $f_e = 10$ and $f_o = 14$. The difference between the observed and the expected frequencies is still 4, but the contribution of this cell to the total chi-square value differs from that of the first case:

$$\text{cell} = \frac{(f_o - f_e)^2}{f_e} = \frac{(14 - 10)^2}{10} = \frac{4^2}{10} = 1.6$$

It should be clear that a small f_e value can have a great influence on the chi-square value. This problem becomes serious when f_e values are less than 5. When f_e is very small, what would otherwise be a minor discrepancy between f_o and f_e results in large chi-square values. The test is too sensitive when f_e values are extremely small. One way to avoid small expected frequencies is to use large samples.

LEARNING CHECK

1. A researcher completes a chi-square test for independence and obtains $\chi^2 = 6.2$ for a sample of $n = 40$ participants.
   a. If the frequency data formed a 2×2 matrix, what is the phi-coefficient for the test?
   b. If the frequency data formed a 3×3 matrix, what is Cramér's V for the test?
2. Explain why a very small value for an expected frequency can distort the results of a chi-square test.

ANSWERS

1. a. $\phi = 0.394$
   b. $V = 0.278$
2. With a very small value for an expected frequency, even a minor discrepancy between the observed frequency and the expected frequency can produce a large number that is added into the chi-square statistic. This inflates the value of chi-square and can distort the outcome of the test.

17.6 SPECIAL APPLICATIONS OF THE CHI-SQUARE TESTS

At the beginning of this chapter, we introduced the chi-square tests as examples of nonparametric tests. Although nonparametric tests serve a function that is uniquely their own, they also can be viewed as alternatives to the common parametric tests that were examined in earlier chapters. In general, nonparametric tests are used as substitutes for parametric tests in situations in which one of the following occurs:

1. The data do not meet the assumptions needed for a standard parametric test.
2. The data consist of nominal or ordinal measurements, so that it is impossible to compute standard descriptive statistics such as the mean and standard deviation.

In this section, we examine some of the relationships between chi-square tests and the parametric procedures for which they may be substituted.

CHI-SQUARE AND THE PEARSON CORRELATION

The chi-square test for independence and the Pearson correlation are both statistical techniques intended to evaluate the relationship between two variables. The type of data obtained in a research study determines which of these two statistical procedures is appropriate. Suppose, for example, that a researcher is interested in the relationship between self-esteem and academic performance for 10-year-old children. If the researcher obtained numerical scores for both variables, then the resulting data would be similar to the values shown in Table 17.11(a) and the researcher could use a Pearson correlation to evaluate the relationship. On the other hand, if both variables are classified into non-numerical categories as in Table 17.11(b), then the data consist of frequencies and the relationship could be evaluated with a chi-square test for independence.

TABLE 17.11

Two possible data structures for research studies examining the relationship between self-esteem and academic performance. In part (a) there are numerical scores for both variables and the data are suitable for a correlation. In part (b) both variables are classified into categories and the data are frequencies suitable for a chi-square test.

(a)

Participant	Self-Esteem *X*	Academic Performance *Y*
A	13	73
B	19	88
C	10	71
D	22	96
E	20	90
F	15	82
.	.	.
.	.	.
.	.	.

(b)

		Level of Self-Esteem			
		High	Medium	Low	
Academic Performance	High	17	32	11	60
	Low	13	43	34	90
		30	75	45	$n = 150$

CHI-SQUARE AND THE INDEPENDENT-MEASURES *t* AND ANOVA

Once again, consider a researcher investigating the relationship between self-esteem and academic performance for 10-year-old children. This time, suppose that the researcher measured academic performance by simply classifying individuals into two categories, high and low, and then obtained a numerical score for each individual's self-esteem. The resulting data would be similar to the scores in Table 17.12(a), and an independent-measures *t* test would be used to evaluate the mean difference between the two groups of scores. Alternatively, the researcher could measure self-esteem by classifying individuals into three categories: high, medium, and low. If a numerical score is then obtained for each individual's academic performance, the resulting data would look like the scores in Table 17.12(b), and an ANOVA would be used to evaluate the mean differences among the three groups. Finally, if both variables are classified into non-numerical categories, then the data would look like the scores shown earlier in Table 17.11(b) and a chi-square test for independence would be used to evaluate the difference between the two academic-performance groups or the differences among the three self-esteem groups.

The point of these examples is that the chi-square test for independence, the Pearson correlation, and tests for mean differences can all be used to evaluate the relationship between two variables. One main distinction among the different statistical procedures is the form of the data. However, another distinction is the fundamental purpose of these different statistics. The chi-square test and the tests for mean differences (*t* and ANOVA) evaluate the *significance* of the relationship; that is, they determine whether the relationship observed in the sample provides enough evidence to conclude that there is a corresponding relationship in the population. You can also evaluate the significance of a Pearson correlation, however, the main purpose of a correlation is to measure the *strength* of the relationship. In particular, squaring the correlation, r^2,

TABLE 17.12

Data appropriate for an independent-measures *t* test or an ANOVA. In part (a), self-esteem scores are obtained for two groups of students differing in level of academic performance. In part (b), academic performance scores are obtained for three groups of students differing in level of self-esteem.

(a) Self-esteem scores for two groups of students.

Academic Performance	
High	Low
17	13
21	15
16	14
24	20
18	17
15	14
19	12
20	19
18	16

(b) Academic performance scores for three groups of students.

Self-esteem		
High	Medium	Low
94	83	80
90	76	72
85	70	81
84	81	71
89	78	77
96	88	70
91	83	78
85	80	72
88	82	75

provides a measure of effect size, describing the proportion of variance in one variable that is accounted for by its relationship with the other variable.

THE MEDIAN TEST FOR INDEPENDENT SAMPLES

The median is the score that divides the population in half, with 50% scoring at or below the median.

The *median test* provides a nonparametric alternative to the independent-measures *t* test (or ANOVA) to determine whether there are significant differences among two or more independent samples. The null hypothesis for the median test states that the different samples come from populations that share a common median (no differences). The alternative hypothesis states that the samples come from populations that are different and do not share a common median.

The logic behind the median test is that whenever several different samples are selected from the same population distribution, roughly half of the scores in each sample should be above the population median and roughly half should be below. That is, all of the separate samples should be distributed around the same median. On the other hand, if the samples come from populations with different medians, then the scores in some samples will be consistently higher and the scores in other samples will be consistently lower.

The first step in conducting the median test is to combine all of the scores from the separate samples and then find the median for the combined group (see Chapter 3, page 83, for instructions for finding the median). Next, a matrix is constructed with a column for each of the separate samples and two rows: one for individuals with scores above the median and one for individuals with scores below the median. Finally, for each sample, count how many individuals scored above the combined median and how many scored below. These values are the observed frequencies that are entered in the matrix.

The frequency-distribution matrix is evaluated using a chi-square test for independence. The expected frequencies and a value for chi-square are computed exactly as described in Section 17.3. A significant value for chi-square indicates that the discrepancy between the individual sample distributions is greater than would be expected by chance.

The median test is demonstrated in the following example.

EXAMPLE 17.4

The following data represent self-esteem scores obtained from a sample of $n = 40$ children. The children are then separated into three groups based on their level of academic performance (high, medium, low). The median test evaluates whether there is a significant relationship between self-esteem and level of academic performance.

Self-Esteem Scores for Children at Three Levels of Academic Performance

High		Medium				Low	
22	14	22	13	24	20	11	19
19	18	18	22	10	16	13	15
12	21	19	15	14	19	20	16
20	18	11	18	11	10	10	18
23	20	12	19	15	12	15	11

The median for the combined group of $n = 40$ scores is $X = 17$ (exactly 20 scores are above this value and 20 are below). For the high performers, 8 out of 10 scores are above the median. For the medium performers, 9 out of 20 are above the median, and for the low performers, only 3 out of 10 are above the median. These observed frequencies are shown in the following matrix:

	Academic Performance		
	High	Medium	Low
Above Median	8	9	3
Below Median	2	11	7

The expected frequencies for this test are as follows:

	Academic Performance		
	High	Medium	Low
Above Median	5	10	5
Below Median	5	10	5

The chi-square statistic is

$$\chi^2 = \frac{9}{5} + \frac{9}{5} + \frac{1}{10} + \frac{1}{10} + \frac{4}{5} + \frac{4}{5} = 5.40$$

With $df = 2$ and $\alpha = .05$, the critical value for chi-square is 5.99. The obtained chi-square of 5.40 does not fall in the critical region, so we would fail to reject the null hypothesis. These data do not provide sufficient evidence to conclude that there are significant differences among the self-esteem distributions for these three groups of students.

A few words of caution are in order concerning the interpretation of the median test. First, the median test is *not* a test for mean differences. Remember: The mean for a distribution can be strongly affected by a few extreme scores. Therefore, the mean and the median for a distribution are not necessarily the same, and they may not even be related. The results from a median test *cannot* be interpreted as indicating that there is (or is not) a difference between means.

Second, you may have noted that the median test does not directly compare the median from one sample with the median from another. Thus, the median test is not a test for significant differences between medians. Instead, this test compares the distribution of scores for one sample to the distribution for another sample. If the samples are distributed evenly around a common point (the group median), then you can conclude that there is no significant difference. On the other hand, finding a significant difference simply indicates that the samples are not distributed evenly around the common median. Thus, the best interpretation of a significant result is that there is a *difference in the distributions* of the samples.

SUMMARY

1. Chi-square tests are nonparametric techniques that test hypotheses about the form of the entire frequency distribution. Two types of chi-square tests are the test for goodness of fit and the test for independence. The data for these tests consist of the frequency or number of individuals who are located in each category.
2. The test for goodness of fit compares the frequency distribution for a sample to the population distribution that is predicted by H_0. The test determines how well the observed frequencies (sample data) fit the expected frequencies (data predicted by H_0).
3. The expected frequencies for the goodness-of-fit test are determined by

$$\text{expected frequency} = f_e = pn$$

   where p is the hypothesized proportion (according to H_0) of observations falling into a category and n is the size of the sample.
4. The chi-square statistic is computed by

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

   where f_o is the observed frequency for a particular category and f_e is the expected frequency for that category. Large values for χ^2 indicate that there is a large discrepancy between the observed (f_o) and the expected (f_e) frequencies and may warrant rejection of the null hypothesis.
5. Degrees of freedom for the test for goodness of fit are

$$df = C - 1$$

   where C is the number of categories in the variable. Degrees of freedom measure the number of categories for which f_e values can be freely chosen. As can be seen from the formula, all but the last f_e value to be determined are free to vary.
6. The chi-square distribution is positively skewed and begins at the value of zero. Its exact shape is determined by degrees of freedom.
7. The test for independence is used to assess the relationship between two variables. The null hypothesis states that the two variables in question are independent of each other. That is, the frequency distribution for one variable does not depend on the categories of the second variable. On the other hand, if a relationship does exist, then the form of the distribution for one variable depends on the categories of the other variable.
8. For the test for independence, the expected frequencies for H_0 can be directly calculated from the marginal frequency totals,

$$f_e = \frac{f_c f_r}{n}$$

   where f_c is the total column frequency and f_r is the total row frequency for the cell in question.
9. Degrees of freedom for the test for independence are computed by

$$df = (R - 1)(C - 1)$$

   where R is the number of row categories and C is the number of column categories.
10. For the test for independence, a large chi-square value means there is a large discrepancy between the f_o and f_e values. Rejecting H_0 in this test provides support for a relationship between the two variables.
11. Both chi-square tests (for goodness of fit and independence) are based on the assumption that each observation is independent of the others. That is, each observed frequency reflects a different individual, and no individual can produce a response that would be

classified in more than one category or more than one frequency in a single category.

12. The chi-square statistic is distorted when f_e values are very small. Chi-square tests, therefore, should not be performed when the expected frequency of any cell is less than 5.

13. The effect size for a chi-square test for independence is measured by computing a phi-coefficient for data that form a 2×2 matrix or computing Cramér's V for a matrix that is larger than 2×2.

$$\text{phi} = \sqrt{\frac{\chi^2}{n}} \qquad \text{Cramér's } V = \sqrt{\frac{\chi^2}{n(df^*)}}$$

where df^* is the smaller of $(R - 1)$ and $(C - 1)$. Both phi and Cramér's V are evaluated using the criteria in Table 17.10.

KEY TERMS

parametric test (593)
nonparametric test (593)
chi-square test for goodness-of-fit (594)
observed frequencies (597)
expected frequencies (597)
chi-square statistic (598)
chi-square distribution (599)
chi-square test for independence (605)
phi-coefficient (613)
Cramér's V (614)
median test (618)

RESOURCES

Book Companion Website: www.cengage.com/psychology/gravetter
You can find a tutorial quiz and other learning exercises for Chapter 17 on the book companion website. The website also provides access to a workshop entitled *Chi-Square* that reviews the chi-square tests presented in this chapter.

aplia

Improve your understanding of statistics with Aplia's auto-graded problem sets and immediate, detailed explanations for every question. To learn more, visit **www.aplia.com/statistics.**

CENGAGEbrain.com

Log in to CengageBrain to access the resources your instructor requires. For this book, you can access:

Psychology CourseMate brings course concepts to life with interactive learning, study, and exam preparation tools that support the printed textbook. A textbook-specific website, Psychology CourseMate includes an integrated interactive eBook and other interactive learning tools including quizzes, flashcards, and more.

Visit www.cengagebrain.com to access your account and purchase materials.

SPSS

General instructions for using SPSS are presented in Appendix D. Following are detailed instructions for using SPSS to perform **The Chi-Square Tests for Goodness of Fit and for Independence** that are presented in this chapter.

The Chi-Square Test for Goodness of Fit

Data Entry

1. Enter the set of observed frequencies in the first column of the SPSS data editor. If there are four categories, for example, enter the four observed frequencies.
2. In the second column, enter the numbers 1, 2, 3, and so on, so that there is a number beside each of the observed frequencies in the first column.

Data Analysis

1. Click **Data** on the tool bar at the top of the page and select **weight cases** at the bottom of the list.
2. Click the **Weight cases by** circle, then highlight the label for the column containing the observed frequencies (VAR00001) on the left and move it into the **Frequency Variable** box by clicking on the arrow.
3. Click **OK.**
4. Click **Analyze** on the tool bar, select **Nonparametric Tests,** and click on **Chi-Square.**
5. Highlight the label for the column containing the digits 1, 2, 3, and move it into the Test Variables box by clicking on the arrow.
6. To specify the expected frequencies, you can either use the **all categories equal** option, which automatically computes expected frequencies, or you can enter your own values. To enter your own expected frequencies, click on the **values** option, and, one by one, enter the expected frequencies into the small box and click **Add** to add each new value to the bottom of the list.
7. Click **OK.**

SPSS Output

The program produces a table showing the complete set of observed and expected frequencies. A second table provides the value for the chi-square statistic, the degrees of freedom, and the level of significance (the *p* value, or alpha level, for the test).

The Chi-Square Test for Independence

Data Entry

1. Enter the complete set of observed frequencies in one column of the SPSS data editor (VAR00001).
2. In a second column, enter a number (1, 2, 3, etc.) that identifies the row corresponding to each observed frequency. For example, enter a 1 beside each observed frequency that came from the first row.
3. In a third column, enter a number (1, 2, 3, etc.) that identifies the column corresponding to each observed frequency. Each value from the first column gets a 1, and so on.

Data Analysis

1. Click **Data** on the tool bar at the top of the page and select **weight cases** at the bottom of the list.
2. Click the **Weight cases by** circle, then highlight the label for the column containing the observed frequencies (VAR00001) on the left and move it into the **Frequency Variable** box by clicking on the arrow.

3. Click **OK.**
4. Click **Analyze** on the tool bar at the top of the page, select **Descriptive Statistics,** and click on **Crosstabs.**
5. Highlight the label for the column containing the rows (VAR00002) and move it into the **Rows** box by clicking on the arrow.
6. Highlight the label for the column containing the columns (VAR00003) and move it into the **Columns** box by clicking on the arrow.
7. Click on **Statistics,** select **Chi-Square,** and click **Continue.**
8. Click **OK.**

SPSS Output

We used SPSS to conduct the chi-square test for independence for the data in Example 17.2, examining the relationship between gender and willingness to use mental health services, and the output is shown in Figure 17.5. The first table in the output simply lists the variables and is not shown in the figure. The **Crosstabulation** table simply shows the matrix of observed frequencies. The final table, labeled **Chi-Square Tests,** reports the results. Focus on the top row, the **Pearson Chi-Square,** which reports the calculated chi-square value, the degrees of freedom, and the level of significance (the *p* value, or the alpha level, for the test).

VAR00002*VAR00003 Crosstabulation

Count

		VAR00003			
		1.00	2.00	3.00	Total
VAR00002	1.00	17	32	11	60
	2.00	13	43	34	90
Total		30	75	45	150

Chi-Square Tests

	Value	df	Asymp.Sig. (2-sided)
Pearson Chi-Square	8.231[a]	2	.016
Likelihood Ratio	8.443	2	.015
Linear by Linear Association	8.109	1	.004
N of Valid Cases	150		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 12.00.

FIGURE 17.5

The SPSS output for the chi-square test for independence in Example 17.2.

FOCUS ON PROBLEM SOLVING

1. The expected frequencies that you calculate must satisfy the constraints of the sample. For the goodness-of-fit test, $\Sigma f_e = \Sigma f_o = n$. For the test for independence, the row totals and column totals for the expected frequencies should be identical to the corresponding totals for the observed frequencies.
2. It is entirely possible to have fractional (decimal) values for expected frequencies. Observed frequencies, however, are always whole numbers.
3. Whenever $df = 1$, the difference between observed and expected frequencies ($f_o = f_e$) is identical (the same value) for all cells. This makes the calculation of chi-square easier.
4. Although you are advised to compute expected frequencies for all categories (or cells), you should realize that it is not essential to calculate all f_e values separately. Remember that df for chi-square identifies the number of f_e values that are free to vary. Once you have calculated that number of f_e values, the remaining f_e values are determined. You can get these remaining values by subtracting the calculated f_e values from their corresponding row or column totals.
5. Remember that, unlike previous statistical tests, the degrees of freedom (df) for a chi-square test are *not* determined by the sample size (n). Be careful!

DEMONSTRATION 17.1

TEST FOR INDEPENDENCE

A manufacturer of watches would like to examine preferences for digital versus analog watches. A sample of $n = 200$ people is selected, and these individuals are classified by age and preference. The manufacturer would like to know whether there is a relationship between age and watch preference. The observed frequencies (f_o) are as follows:

	Digital	Analog	Undecided	Totals
Younger than 30	90	40	10	140
30 or Older	10	40	10	60
Column totals	100	80	20	$n = 200$

STEP 1 State the hypotheses, and select an alpha level.

The null hypothesis states that there is no relationship between the two variables.

H_0: Preference is independent of age. That is, the frequency distribution of preference has the same form for people younger than 30 as for people 30 or older.

The alternative hypothesis states that there is a relationship between the two variables.

H_1: Preference is related to age. That is, the type of watch preferred depends on a person's age.

We set alpha to $\alpha = .05$.

STEP 2 Locate the critical region.

Degrees of freedom for the chi-square test for independence are determined by

$$df = (C - 1)(R - 1)$$

For these data,

$$df = (3 - 1)(2 - 1) = 2(1) = 2$$

For $df = 2$ with $\alpha = .05$, the critical chi-square value is 5.99. Thus, our obtained chi-square must exceed 5.99 to be in the critical region and to reject H_0.

STEP 3 Compute the test statistic. Two calculations are required: finding the expected frequencies and calculating the chi-square statistic.

Expected frequencies, f_e. For the test for independence, the expected frequencies can be found using the column totals (f_c), the row totals (f_r), and the following formula:

$$f_e = \frac{f_c f_r}{n}$$

For people younger than 30, we obtain the following expected frequencies:

$$f_e = \frac{100(140)}{200} = \frac{14{,}000}{200} = 70 \text{ for digital}$$

$$f_e = \frac{80(140)}{200} = \frac{11{,}200}{200} = 56 \text{ for analog}$$

$$f_e = \frac{20(140)}{200} = \frac{2800}{200} = 14 \text{ for undecided}$$

For individuals 30 or older, the expected frequencies are as follows:

$$f_e = \frac{100(60)}{200} = \frac{6000}{200} = 30 \text{ for digital}$$

$$f_e = \frac{80(60)}{200} = \frac{4800}{200} = 24 \text{ for analog}$$

$$f_e = \frac{20(60)}{200} = \frac{1200}{200} = 6 \text{ for undecided}$$

The following table summarizes the expected frequencies:

	Digital	Analog	Undecided
Younger than 30	70	56	14
30 or Older	30	24	6

The chi-square statistic. The chi-square statistic is computed from the formula

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

The following table summarizes the calculations:

Cell	f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
Younger than 30—digital	90	70	20	400	5.71
Younger than 30—analog	40	56	−16	256	4.57
Younger than 30—undecided	10	14	−4	16	1.14
30 or Older—digital	10	30	−20	400	13.33
30 or Older—analog	40	24	16	256	10.67
30 or Older—undecided	10	6	4	16	2.67

Finally, add the values in the last column to get the chi-square statistic.

$$\chi^2 = 5.71 + 4.57 + 1.14 + 13.33 + 10.67 + 2.67$$
$$= 38.09$$

STEP 4 Make a decision about H_0, and state the conclusion.

The chi-square value is in the critical region. Therefore, we reject the null hypothesis. There is a relationship between watch preference and age, $\chi^2(2, n = 200) = 38.09$, $p < .05$.

DEMONSTRATION 17.2

EFFECT SIZE WITH CRAMÉR'S *V*

Because the data matrix is larger than 2 × 2, we compute Cramér's V to measure effect size.

$$\text{Cramér's } V = \sqrt{\frac{\chi^2}{n(df^*)}} = \sqrt{\frac{38.09}{200(1)}} = \sqrt{0.19} = 0.436$$

PROBLEMS

1. Parametric tests (such as *t* or ANOVA) differ from nonparametric tests (such as chi-square) primarily in terms of the assumptions they require and the data they use. Explain these differences.

2. The student population at the state college consists of 55% females and 45% males.

a. The college theater department recently staged a production of a modern musical. A researcher recorded the gender of each student entering the theater and found a total of 385 females and 215 males. Is the gender distribution for theater goers significantly different from the distribution for the general college? Test at the .05 level of significance.

b. The same researcher also recorded the gender of each student watching a men's basketball game in the college gym and found a total of 83 females and

97 males. Is the gender distribution for basketball fans significantly different from the distribution for the general college? Test at the .05 level of significance.

3. A developmental psychologist would like to determine whether infants display any color preferences. A stimulus consisting of four color patches (red, green, blue, and yellow) is projected onto the ceiling above a crib. Infants are placed in the crib, one at a time, and the psychologist records how much time each infant spends looking at each of the four colors. The color that receives the most attention during a 100-second test period is identified as the preferred color for that infant. The preferred colors for a sample of 60 infants are shown in the following table:

Red	Green	Blue	Yellow
20	12	18	10

a. Do the data indicate any significant preferences among the four colors? Test at the .05 level of significance.

b. Write a sentence demonstrating how the outcome of the hypothesis test would appear in a research report.

4. Data from the department of motor vehicles indicate that 80% of all licensed drivers are older than age 25.

a. In a sample of $n = 60$ people who recently received speeding tickets, 38 were older than 25 years and the other 22 were age 25 or younger. Is the age distribution for this sample significantly different from the distribution for the population of licensed drivers? Use $\alpha = .05$.

b. In a sample of $n = 60$ people who recently received parking tickets, 43 were older than 25 years and the other 17 were age 25 or younger. Is the age distribution for this sample significantly different from the distribution for the population of licensed drivers? Use $\alpha = .05$.

5. To investigate the phenomenon of "home-team advantage," a researcher recorded the outcomes from 64 college football games on one Saturday in October. Of the 64 games, 42 were won by home teams. Does this result provide enough evidence to conclude that home teams win significantly more than would be expected by chance? Assume that winning and losing are equally likely events if there is no home-team advantage. Use $\alpha = .05$.

6. Research has demonstrated that people tend to be attracted to others who are similar to themselves. One study demonstrated that individuals are disproportionately more likely to marry those with surnames that begin with the same last letter as their own (Jones, Pelham, Carvallo, & Mirenberg, 2004). The researchers began by looking at marriage records and recording the surname for each groom and the maiden name of each bride. From these records it is possible to calculate the probability of randomly matching a bride and a groom whose last names begin with the same letter. Suppose that this probability is only 6.5%. Next, a sample of $n = 200$ married couples is selected and the number who shared the same last initial at the time they were married is counted. The resulting observed frequencies are as follows:

Same Initial	Different Initials	
19	181	200

Do these data indicate that the number of couples with the same last initial is significantly different that would be expected if couples were matched randomly? Test with $\alpha = .05$.

7. Suppose that the researcher from the previous problem repeated the study of married couples' initials using twice as many participants and obtaining observed frequencies that exactly double the original values. The resulting data are as follows:

Same Initial	Different Initials	
38	362	400

a. Use a chi-square test to determine whether the number of couples with the same last initial is significantly different than would be expected if couples were matched randomly. Test with $\alpha = .05$.

b. You should find that the data lead to rejecting the null hypothesis. However, in problem 6 the decision was fail to reject. How do you explain the fact that the two samples have the same proportions but lead to different conclusions?

8. A professor in the psychology department would like to determine whether there has been a significant change in grading practices over the years. It is known that the overall grade distribution for the department in 1985 had 14% As, 26% Bs, 31% Cs, 19% Ds, and 10% Fs. A sample of $n = 200$ psychology students from last semester produced the following grade distribution:

A	B	C	D	F
32	61	64	31	12

Do the data indicate a significant change in the grade distribution? Test at the .05 level of significance.

9. Automobile insurance is much more expensive for teenage drivers than for older drivers. To justify this cost difference, insurance companies claim that the younger drivers are much more likely to be involved in costly accidents. To test this claim, a researcher obtains information about registered drivers from the department of motor vehicles (DMV) and selects a sample of $n = 300$ accident reports from the police department. The DMV reports the percentage of registered drivers in each age category as follows: 16% are younger than age 20; 28% are 20 to 29 years old; and 56% are age 30 or older. The number of accident reports for each age group is as follows:

Under age 20	Age 20–29	Age 30 or older
68	92	140

a. Do the data indicate that the distribution of accidents for the three age groups is significantly different from the distribution of drivers? Test with $\alpha = .05$.

b. Write a sentence demonstrating how the outcome of the hypothesis test would appear in a research report.

10. The color red is often associated with anger and male dominance. Based on this observation, Hill and Barton (2005) monitored the outcome of four combat sports (boxing, tae kwan do, Greco-Roman wrestling, and freestyle wrestling) during the 2004 Olympic games and found that participants wearing red outfits won significantly more often than those wearing blue.

a. In 50 wrestling matches involving red versus blue, suppose that the red outfit won 31 times and lost 19 times. Is this result sufficient to conclude that red wins significantly more than would be expected by chance? Test at the .05 level of significance.

b. In 100 matches, suppose red won 62 times and lost 38. Is this sufficient to conclude that red wins significantly more than would be expected by chance? Again, use $\alpha = .05$.

c. Note that the winning percentage for red uniforms in part a is identical to the percentage in part b (31 out of 50 is 62%, and 62 out of 100 is also 62%). Although the two samples have an identical winning percentage, one is significant and the other is not. Explain why the two samples lead to different conclusions.

11. A communications company has developed three new designs for a cell phone. To evaluate consumer response, a sample of 120 college students is selected and each student is given all three phones to use for 1 week. At the end of the week, the students must identify which of the three designs they prefer. The distribution of preference is as follows:

Design 1	Design 2	Design 3
54	38	28

Do the results indicate any significant preferences among the three designs?

12. In problem 11, a researcher asked college students to evaluate three new cell phone designs. However, the researcher suspects that college students may have criteria that are different from those used by older adults. To test this hypothesis, the researcher repeats the study using a sample of $n = 60$ older adults in addition to a sample of $n = 60$ students. The distribution of preference is as follows:

	Design 1	Design 2	Design 3	
Student	27	20	13	60
Older Adult	21	34	5	60
	48	54	18	

Do the data indicate that the distribution of preferences for older adults is significantly different from the distribution for college students? Test with $\alpha = .05$.

13. Research suggests that romantic background music increases the likelihood that a woman will give her phone number to a man she has just met (Guéguen & Jacoby, 2010). In the study, women spent time in a waiting room with background music playing. In one condition, the music was a popular love song and for the other condition the music was a neutral song. The participant was then moved to another room in which she was instructed to discuss two food products with a young man. The men were part of the study and were selected because they had been rated as average in attractiveness. The experimenter returned to end the study and asked the pair to wait alone for a few minutes. During this time, the man used a scripted line to ask the woman for her phone number. The following table presents data similar to those obtained in the study, showing the number of women who did or did not give their numbers for each music condition.

	Phone Number	No Number	
Romantic Music	21	19	40
Neutral Music	9	31	40
	30	50	

Is there a significant difference between the two types of music? Test with $\alpha = .05$

14. Mulvihill, Obuseh, and Caldwell (2008) conducted a survey evaluating healthcare providers' perception of a new state children's insurance program. One question

asked the providers whether they viewed the reimbursement from the new insurance as higher, lower, or the same as private insurance. Another question assessed the providers' overall satisfaction with the new insurance. The following table presents observed frequencies similar to the study results.

	Satisfied	Not Satisfied	
Less Reimbursement	46	54	100
Same or More Reimbursement	42	18	60
	88	72	

Do the results indicate that the providers' satisfaction of the new program is related to their perception of the reimbursement rates? Test with $\alpha = .05$.

15. A local county is considering a budget proposal that would allocate extra funding toward the renovation of city parks. A survey is conducted to measure public opinion concerning the proposal. A total of 150 individuals respond to the survey: 50 who live within the city limits and 100 from the surrounding suburbs. The frequency distribution is as follows:

	Opinion: Favor	Opinion: Oppose	
City	35	15	50
Suburb	55	45	100
	90	60	

a. Is there a significant difference in the distribution of opinions for city residents compared to those in the suburbs? Test at the .05 level of significance.

b. The relationship between home location and opinion can also be evaluated using the phi-coefficient. If the phi-coefficient were computed for these data, what value would be obtained for phi?

16. The data from problem 15 show no significant difference between the opinions for city residents and those who live in the suburbs. To construct the following data, we simply doubled the sample size from problem 15 so that all of the individual frequencies are twice as big. Notice that the sample *proportions* have not changed.

	Opinion: Favor	Opinion: Oppose	
City	70	30	100
Suburb	110	90	200
	180	120	

a. Test for a significant difference between the city distribution and the suburb distribution using $\alpha = .05$. How does the decision compare with the decision in problem 14? You should find that a larger sample increases the likelihood of a significant result.

b. Compute the phi-coefficient for these data and compare it with the result from problem 15. You should find that the sample size has no effect on the strength of the relationship.

17. In the Preview for this chapter, we discussed a study investigating the relationship between memory for eyewitnesses and the questions they are asked (Loftus & Palmer, 1974). In the study, participants watched a film of an automobile accident and then were questioned about the accident. One group was asked how fast the cars were going when they "smashed into" each other. A second group was asked about the speed when the cars "hit" each other, and a third group was not asked any question about the speed of the cars. A week later, the participants returned to answer additional questions about the accident, including whether they recalled seeing any broken glass. Although there was no broken glass in the film, several students claimed to remember seeing it. The following table shows the frequency distribution of responses for each group.

Verb Used to Ask About the Speed of the Cars	Response to the Question "Did You See Any Broken Glass?": Yes	No
Smashed into	16	34
Hit	7	43
Control (not asked)	6	44

a. Does the proportion of participants who claim to remember broken glass differ significantly from group to group? Test with $\alpha = .05$.

b. Compute Cramérs *V* to measure the size of the treatment effect.

c. Describe how the phrasing of the question influenced the participants' memories.

d. Write a sentence demonstrating how the outcome of the hypothesis test and the measure of effect size would be reported in a journal article.

18. In a study investigating freshman weight gain, the researchers also looked at gender differences in weight (Kasparek, Corwin, Valois, Sargent, & Morris, 2008). Using self-reported heights and weights, they computed the body mass index (BMI) for each student. Based on the BMI scores, the students were classified as either desirable weight or overweight. When the students were further classified by gender,

the researchers found results similar to the frequencies in the following table.

	Desirable Weight	Overweight
Males	74	46
Females	62	18

a. Do the data indicate that the proportion of overweight men is significantly different from the proportion of overweight women? Test with $\alpha = .05$.
b. Compute the phi-coefficient to measure the strength of the relationship.
c. Write a sentence demonstrating how the outcome of the hypothesis test and the measure of effect size would be reported in a journal article.

19. Research results suggest that IQ scores for boys are more variable than IQ scores for girls (Arden & Plomin, 2006). A typical study looking at 10-year-old children classifies participants by gender and by low, average, or high IQ. Following are hypothetical data representing the research results. Do the data indicate a significant difference between the frequency distributions for males and females? Test at the .05 level of significance and describe the difference.

	IQ			
	Low	Average	High	
Boys	18	42	20	80
Girls	12	54	14	80
				$n = 160$

20. Gender differences in dream content are well documented (see Winget & Kramer, 1979). Suppose a researcher studies aggression content in the dreams of men and women. Each participant reports his or her most recent dream. Then each dream is judged by a panel of experts to have low, medium, or high aggression content. The observed frequencies are shown in the following matrix:

		Aggression Content		
		Low	Medium	High
Gender	Female	18	4	2
	Male	4	17	15

Is there a relationship between gender and the aggression content of dreams? Test with $\alpha = .01$.

21. In a study similar to one conducted by Fallon and Rozin (1985), a psychologist prepared a set of silhouettes showing different female body shapes ranging from somewhat thin to somewhat heavy and asked a group of women to indicate which body figure they thought men would consider the most attractive. Then a group of men were shown the same set of profiles and asked which image they considered the most attractive. The following hypothetical data show the number of individuals who selected each of the four body image profiles.
a. Do the data indicate a significant difference between the actual preferences for the men and the preferences predicted by the women? Test at the .05 level of significance.
b. Compute the phi-coefficient to measure the strength of the relationship.

	Body Image Profiles				
	Somewhat Thin	Slightly Thin	Slightly Heavy	Somewhat Heavy	
Women	29	25	18	8	80
Men	11	15	22	12	60
	40	40	40	20	

22. A recent study indicates that people tend to select video game avatars with characteristics similar to those of their creators (Bélisle & Onur, 2010). Participants who had created avatars for a virtual community game completed a questionnaire about their personalities. An independent group of viewers examined the avatars and recorded their impressions of the avatars. One personality characteristic considered was introverted/extroverted. The following frequency distribution of personalities for participants and the avatars they created.

	Participant Personality		
	Introverted	Extroverted	
Introverted Avatar	22	23	45
Extroverted Avatar	16	39	55
	38	62	

a. Is there a significant relationship between the personalities of the participants and the personalities of their avatars? Test with $\alpha = .05$.
b. Compute the phi-coefficient to measure the size of the effect.

23. Research indicates that people who volunteer to participate in research studies tend to have higher intelligence than nonvolunteers. To test this

phenomenon, a researcher obtains a sample of 200 high school students. The students are given a description of a psychological research study and asked whether they would volunteer to participate. The researcher also obtains an IQ score for each student and classifies the students into high, medium, and low IQ groups. Do the following data indicate a significant relationship between IQ and volunteering? Test at the .05 level of significance.

	IQ			
	High	Medium	Low	
Volunteer	43	73	34	150
Not Volunteer	7	27	16	50
	50	100	50	

24. Cialdini, Reno, and Kallgren (1990) examined how people conform to norms concerning littering. The researchers wanted to determine whether a person's tendency to litter depended on the amount of litter already in the area. People were handed a handbill as they entered an amusement park. The entrance area had already been prepared with either no litter, a small amount of litter, or a lot of litter lying on the ground. The people were observed to determine whether they dropped their handbills. The frequency data are as follows:

	Amount of Litter		
	None	Small Amount	Large Amount
Littering	17	28	49
Not Littering	73	62	41

a. Do the data indicate that people's tendency to litter depends on the amount of litter already on the ground? That is, is there a significant relationship between littering and the amount of existing litter? Test at the .05 level of significance.

b. Compute Cramér's V to measure the size of the treatment effect.

25. Although the phenomenon is not well understood, it appears that people born during the winter months are slightly more likely to develop schizophrenia than people born at other times (Bradbury & Miller, 1985). The following hypothetical data represent a sample of 50 individuals diagnosed with schizophrenia and a sample of 100 people with no psychotic diagnosis. Each individual is also classified according to season in which he or she was born. Do the data indicate a significant relationship between schizophrenia and the season of birth? Test at the .05 level of significance.

	Season of Birth				
	Summer	Fall	Winter	Spring	
No Disorder	26	24	22	28	100
Schizophrenia	9	11	18	12	50
	35	35	40	40	

This page intentionally left blank