Chi-Square Tests and Strategies When Population Distributions Are Not Normal

Chapter Outline

- Chi-Square Tests
- The Chi-Square Statistic and the Chi-Square Test for Goodness of Fit
- The Chi-Square Test for Independence
- Assumptions for the Chi-Square Tests
- Effect Size and Power for the Chi-Square Tests for Independence
- Strategies for Hypothesis Testing When Population Distributions Are Not Normal
- Data Transformations
- Rank-Order Tests
- Comparison of Methods
- Chi-Square Tests, Data Transformations, and Rank-Order Tests in Research Articles
- Learning Aids

  Summary

  Key Terms

  Example Worked-Out Problems

  Practice Problems

  Using SPSS

The hypothesis-testing procedures you have learned in the chapter (the *t* test and the analysis of variance) are very versatile, but there are certain research situations in which these methods cannot be used. One such situation is hypothesis testing for variables whose values are categories, such as a person's region of the country, religious preference, or hair color.

The *t* test and the analysis of variance all require that the measured variable have scores that are quantitative, such as a rating on a 7-point scale or number of years served as mayor. Another research situation in which the ordinary *t* test and analysis of variance do not apply is when the populations do not clearly follow a normal curve.

TIP FOR SUCCESSS

This chapter assumes you have a solid command of hypothesis testing and are familiar with the *t* test and analysis of variance, and in particular the assumptions required for their use.

From Chapter 11 of *Statistics for the Behavioral & Social Sciences: A Brief Course*, Fifth Edition. Arthur Aron, Elliot J. Coups, Elaine N. Aron. Copyright © 2011 by Pearson Education, Inc. Published by Pearson Prentice Hall. All rights reserved.

This chapter examines hypothesis testing in these two situations in which the ordinary hypothesis-testing procedures cannot be used properly. The first half of the chapter focuses on chi-square tests. **Chi-square tests** are used when the variable of interest is a *nominal variable* (that is, a variable with values that are categories). (Chi is the Greek letter χ and is pronounced *ki,* rhyming with *high* and *pie.*) Therefore, the scores in this situation represent *frequencies:* that is, how many people or observations fall into different categories. The chi-square test was originally developed by Karl Pearson (see Box 1) and is sometimes called the *Pearson chi-square.* The second half of the chapter focuses on strategies for hypothesis testing when you cannot assume that the population distributions are even roughly normal.

Chi-Square Tests

Consider an example. Black, Marola, Littman, Chrisler, and Neace (2009) were interested in the extent to which cereal boxes are more likely to show male as opposed to female characters. Previous research has shown that male characters outnumber female characters in various forms of media targeted toward children, including books, video games, and television. The researchers noted that "cereal boxes . . . depict a plethora of characters used to market the products to both children and adults" and that "... these boxes are on the kitchen table each morning as people eat their breakfast, and may be more prevalent in households than some other forms of media that have been studied previously. As such, the gender of the characters and manner in which they are depicted on cereal boxes may contribute to people's gender schemas" (pp. 882–883). In order to test their hypothesis that male characters would appear more often than female characters on cereal boxes, the researchers had student research assistants code the gender of the characters on every cereal box in a large grocery superstore in the northeastern United States. The researchers described the coding process as follows: "[F]or gender, coders relied on cues such as clothing (e.g., skirts), hairstyle (e.g., pony tails with ribbons), facial features (e.g., mustache), and name (e.g., Tony the Tiger)" (p. 885). Of the 217 cereal boxes that the students coded, 166 had one or more characters on the box. Of the 1,386 characters whose gender was determined, 996 were male and 390 were female characters. In terms of percentages, there were 72% male and 28% female characters.

Suppose the characters were equally likely to be male or female. If that were the case, then about 693 (half of the 1,386) characters should have been male and another 693 should have been female. This information is laid out in the "Observed Frequency" and "Expected Frequency" columns of Table 1. The Observed Frequency column shows the breakdown of character genders actually *observed.* The Expected Frequency column shows the breakdown you would *expect* if the genders had been exactly equally

chi-square test Hypothesis-testing procedure used when the variables of interest are nominal variables.

Table 1 Observed and Expected Frequencies for the Gender of Characters on Cereal Boxes (Data from Black et al., 2009)

Gender	Observed Frequency (O)	Expected Frequency (E)	Difference ($O - E$)	Difference Squared ($O - E)^2$	Difference Squared Weighted by Expected Frequency $(O - E)^2/E$
Male	996	693	303	91,809	132.48
Female	390	693	−303	91,809	132.48

BOX 1 Karl Pearson: Inventor of Chi-Square and Center of Controversy

Topham/The Image Works

Karl Pearson, sometimes hailed as the founder of the science of statistics, was born in 1857. Both his virtues and vices are revealed in what he reported to his colleague Julia Bell as his earliest memory: he was sitting in his highchair, sucking his thumb, when he was told to stop or his thumb would wither away. Pearson silently thought, "I can't see that the thumb I suck is any smaller than the other. I wonder if she could be lying to me." Here we see Pearson's faith in himself and in observational evidence, as well as his rejection of authority. We also see his tendency to doubt the character of people with whom he disagreed.

Pearson studied mathematics at Cambridge. Soon after he arrived, he requested to be excused from compulsory chapel. As soon as his request was granted, however, he appeared in chapel. The dean summoned him for an explanation, and Pearson declared that he had asked to be excused not from chapel "but from *compulsory* chapel."

After graduation, Pearson studied in Germany, becoming a socialist and a self-described "free-thinker." Returning to England, he changed his name from Carl to Karl and wrote an attack on Christianity under a pen name. In 1885 he founded a Men and Women's Club to promote equality between the sexes. In 1892 he published a book titled *The Grammar of Science* that subsequently influenced Albert Einstein's theories of relativity.

Most of Pearson's research from 1893 to 1901 focused on the laws of heredity, but he needed better statistical methods, leading to his most famous contribution, the chi-square test. Pearson also invented the method of computing correlation used today and coined the terms *histogram, skew,* and *spurious correlation.* When he felt that biology journals failed to appreciate his work, he founded the famous journal *Biometrika.*

Unfortunately, Pearson was a great fan of eugenics, and his work was later used by the Nazis as justification for their treatment of the Jews. Toward the end of his life, however, he wrote a paper using clear logic and data on Jews and Gentiles from all over the world to demonstrate that the Nazis' ideas were sheer nonsense.

Indeed, throughout his life, Pearson's strong opinions created a long list of enemies, especially as other, younger statisticians passed him by, while he refused to publish their work in *Biometrika.* William S. Gosset was one of his friends. Sir Ronald Fisher was one of Pearson's worst enemies. The kindly Gosset was always trying to smooth matters between them. In 1933, Pearson finally retired, and Fisher, of all persons, took over his chair, the Galton Professorship of Eugenics at University College in London. In 1936, the two entered into their bitterest argument yet; Pearson died the same year.

For more information about Pearson, see http://en.wikipedia.org/wiki/Karl_Pearson and http://human-nature.com/nibbs/03/kpearson.html.

Sources: Peters (1987), Salsburg (2001), Stigler (1986), Tankard (1984), Wright (2009).

likely. (Note that it won't always be the case that you expect an *equal breakdown* across the categories. In other situations, the expected frequency for each category may be based on theory, or on a distribution in another study or circumstance.)

Clearly, there is a discrepancy between what was actually observed and the breakdown you would expect if male and female characters were equally likely. The question is this: Should you assume that this discrepancy is no more than what we would expect just by chance for a sample of this size? Suppose that characters on all cereal boxes (that is, the entire population of cereal boxes, not just the sample the students saw at that supermarket) are equally likely to be male or female. In that case, you would not expect a perfectly equal gender split for the sample of cereal boxes examined from any single store. But if the breakdown in the sample of cereal boxes is a long way from equal, you would doubt that the gender split in the full population of cereal boxes really is equal. In other words, we are in a hypothesis-testing situation, much like the ones we have been considering all along. But with a big difference too.

The scores have all been *numerical values* on some dimension, such as a score on a standard achievement test, length of time in a relationship, an employer's rating of an employee's job effectiveness on a 9-point scale, and so forth; often we figured means of these numbers. By contrast, gender of characters on cereal boxes is an example of a *nominal variable* (or a *categorical variable*). A nominal variable is one in which the information is the number of people or observations in each category. Therefore, the numbers associated with nominal variables are frequencies (and not means at all); the frequency tells you how many people or observations fall into each category of the variable. We use the term *nominal variables* because the different categories or levels of the variable have names instead of numbers. Hypothesis testing with nominal variables uses what are called *chi-square tests.*

The Chi-Square Statistic and the Chi-Square Test for Goodness of Fit

The basic idea of any chi-square test is that you compare how well an *observed breakdown* of people or observations over various categories fits some *expected breakdown* (such as an equal breakdown). In this chapter, you will learn about two types of chi-square tests. First, you will learn about the **chi-square test for goodness of fit,** which is a chi-square test involving levels of a *single nominal variable.* Later in the chapter, you will learn about the **chi-square test for independence,** which is used when there are *two nominal variables,* each with several categories.

In terms of the example of characters on cereal boxes—in which there is a single nominal variable with two categories (male and female)—you are comparing the observed breakdown of 996 and 390 to the expected breakdown of about 693 for each gender. A breakdown of numbers expected in each category is actually a frequency distribution. Thus, a chi-square test is more formally described as comparing an **observed frequency** distribution to an **expected frequency** distribution. *Here is the key idea: What this hypothesis testing involves is first figuring a number for the amount of mismatch between the observed frequencies and the expected frequencies and then seeing whether that number indicates a greater mismatch than you would expect by chance.* This gives an idea as to how the chi-square test for *goodness of fit* came to have that name: The test shows how well an observed frequency distribution fits an expected (or predicted) frequency distribution.

chi-square test for goodness of fit Hypothesis-testing procedure that examines how well an observed frequency distribution of a single nominal variable fits some expected pattern of frequencies.

chi-square test for independence Hypothesis-testing procedure that examines whether the distribution of frequencies over the categories of one nominal variable are unrelated to (independent of) the distribution of frequencies over the categories of a second nominal variable.

observed frequency In a chi-square test, number of individuals actually found in the study to be in a category or cell.

expected frequency In a chi-square test, number of people in a category or cell expected if the null hypothesis were true.

Let's start with how you would come up with that mismatch number for the observed versus expected frequencies. The mismatch between observed and expected for any one category is just the observed frequency minus the expected frequency. For example, consider again the Black et al. (2009) study. For male characters, the observed frequency of 996 is 303 more than the expected frequency of 693. For female characters, the difference is −303 (that is, $390 - 693 = -303$). These differences are shown in the Difference column of Table 1.

You do not use these differences directly. One reason is that some differences are positive and some are negative. Thus, they would cancel each other out. To get around this, you square each difference. In our example, the squared difference for male characters is 303 squared, or 91,809. For female characters it is −303 squared, which is also 91,809. These squared differences are shown in the Difference Squared column of Table 1.

In the Black et al. (2009) example, the expected frequencies are the same in each category. But in other research situations, expected frequencies for the different categories may not be the same. A particular amount of difference between observed and expected has a different importance according to the size of the expected frequency. For example, a difference of eight people between observed and expected is a much bigger mismatch if the expected frequency is 10 than if the expected frequency is 1,000. If the expected frequency is 10, a difference of 8 would mean that the observed frequency was 18 or 2, frequencies that are dramatically different from 10. But if the expected frequency is 1,000, a difference of 8 is only a slight mismatch. This would mean that the observed frequency was 1,008 or 992, frequencies that are only slightly different from 1,000.

How can you adjust the mismatch (the squared difference) between observed and expected for a particular category? What you need to do is adjust or weight the mismatch to take into account the expected frequency for that category. You can do this by dividing your squared difference for a category by the expected frequency for that category. Thus, if the expected frequency for a particular category is 10, you divide the squared difference by 10. If the expected frequency for the category is 1,000, you divide the squared difference by 1,000. In this way, you weight each squared difference by the expected frequency. This weighting puts the squared difference onto a more appropriate scale of comparison.

Let's return to our example. For male characters, you would weight the mismatch by dividing the squared difference of 91,809 by 693, giving 132.48. For female characters, dividing 91,809 by 693 again gives 132.48. These adjusted mismatches (squared differences weighted by expected frequencies) are shown in the right-most column of Table 1.

What remains is to get an overall number for the mismatch between observed and expected frequencies. This final step is done by adding up the mismatch for all the categories. That is, you take the result of the squared difference divided by the expected frequency for the first category, add the result of the squared difference divided by the expected frequency for the second category, and so on. In the Black et al. (2009) example, there are only two categories and this would be 132.48 plus 132.48, for a total of 264.96. This final number (the sum of the weighted squared differences) is an overall indication of the amount of mismatch between the expected and observed frequencies. It is called the **chi-square statistic.** In terms of a formula,

$$\chi^2 = \sum \frac{(O - E)^2}{E} \qquad (1)$$

TIP FOR SUCCESS

Notice in the chi-square formula that, for each category, you first divide the squared difference between observed and expected frequencies by the expected frequency, and then you sum the resulting values for all the categories. This is a slightly different procedure than you are used to from this chapter (in which you often first summed a series of squared values in the numerator and then divided by a denominator value), so be sure to follow the formula carefully.

Chi-square is the sum, over all the categories, of the squared difference between observed and expected frequencies divided by the expected frequency.

In this formula, χ^2 is the chi-square statistic. Σ is the summation sign, telling you to sum over all the different categories. *O* is the observed frequency for a category (the number of people or observations actually found in that category in the study). *E* is the expected frequency for a category. (In the Black et al. [2009] example, it is based on what we would expect if there were equal numbers in each category.)

Applying the formula to the Black et al. (2009) example,

$$\chi^2 = \Sigma\frac{(O - E)^2}{E} = \frac{(996 - 693)^2}{693} + \frac{(390 - 693)^2}{693} = 264.96$$

Steps for Figuring the Chi-Square Statistic

Here is a summary of what we've said so far in terms of steps:

Ⓐ **Determine the actual, observed frequencies in each category.**
Ⓑ **Determine the expected frequencies in each category.**
Ⓒ **In each category, take observed minus expected frequencies.**

chi-square statistic (χ^2) Statistic that reflects the overall lack of fit between the expected and observed frequencies; the sum, over all the categories, of the squared difference between observed and expected frequencies divided by the expected frequency.

Ⓓ Square each of these differences.
Ⓔ Divide each squared difference by the expected frequency for its category.
Ⓕ Add up the results of Step Ⓔ for all the categories.

The Chi-Square Distribution

Now we turn to the question of whether the chi-square statistic you have figured is a bigger mismatch than you would expect by chance. To answer that, you need to know how likely it is to get chi-square statistics of various sizes by chance. That is, you need the distribution of chi-square statistics that would arise by chance. As long as you have a reasonable number of people in the study, the distribution of the chi-square statistic follows quite closely a known mathematical distribution—the **chi-square distribution.**

The exact shape of the chi-square distribution depends on the degrees of freedom. For a chi-square test, the degrees of freedom are the number of categories that are free to vary, given the totals. In the example of the gender of characters on cereal boxes, there are two categories (male and female). If you know the total number of cereal boxes and you know the number in any one category, you could figure out the number in the second category—so only one category is free to vary. That is, in a study like this one (which uses a chi-square test for goodness of fit), if there are two categories, there is one degree of freedom. In terms of a formula,

The degrees of freedom for the chi-square test for goodness of fit are the number of categories minus 1.

$$df = N_{\text{Categories}} - 1 \tag{2}$$

Chi-square distributions for several different degrees of freedom are shown in Figure 1. Notice that the distributions are all skewed to the right. This is because the chi-square statistic cannot be less than 0 but can have very high values. (Chi-square must be positive because it is figured by adding a group of fractions, in each of which the numerator and denominator both have to be positive. The numerator has to be positive because it is squared. The denominator has to be positive because the number of people expected in a given category can't be a negative number; you can't expect less than no one!)

The Chi-Square Table

chi-square distribution Mathematically defined curve used as the comparison distribution in chi-square tests; the distribution of the chi-square statistic.

What matters most about the chi-square distribution for hypothesis testing is the cutoff for a chi-square to be extreme enough to reject the null hypothesis. For example, suppose you are using the .05 significance level. In that situation, you want to know the point on the chi-square distribution where 5% of the chi-square statistics are above

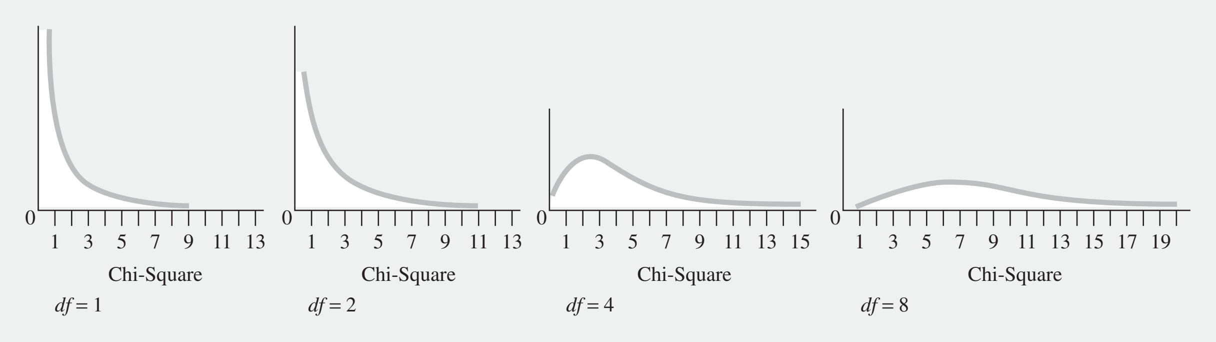


Figure 1 Examples of chi-square distributions for different degrees of freedom.

that point. A **chi-square table** gives the cutoff chi-square for different significance levels and various degrees of freedom. Table 2 shows a portion of a chi-square table. For our example, where there is one degree of freedom, the table shows that the cutoff chi-square for the .05 level is 3.841. (Of course, as with other hypothesis-testing procedures, when you carry out a chi-square using a statistics program like SPSS, you will be given an exact probability level, which you then check to see if it is less than the cutoff you set for the study, such as .05 or .01.)

Table 2 Portion of a Chi-Square Table (with Cutoff Value Highlighted for the Black et al. Example)

	Significance Level		
df	.10	.05	.01
1	2.706	3.841	6.635
2	4.605	5.992	9.211
3	6.252	7.815	11.345
4	7.780	9.488	13.277
5	9.237	071	15.087

Note: Full table is Table A-4 in the Appendix.

The Chi-Square Test for Goodness of Fit

In the Black et al. (2009) example, we figured a chi-square of 264.96. This is clearly larger than the chi-square cutoff for this example (using the .05 significance level) of 3.841 (see Figure 2). Thus, the researchers rejected the null hypothesis. That is, they rejected as too unlikely that the mismatch they observed (between the observed and expected frequencies) could have come about if the entire population of characters on cereal boxes had an equal number of male and female characters. It seemed more reasonable to conclude that there truly were different proportions of male and female characters on cereal boxes.

We have just carried out a full hypothesis-testing procedure for the Black et al. (2009) example. This example involved differing numbers of observations at two levels of a particular nominal variable (the gender of characters on cereal boxes). As we mentioned earlier, this kind of chi-square test involving levels of a single nominal variable is called a *chi-square test for goodness of fit.*

chi-square table Table of cutoff scores on the chi-square distribution for various degrees of freedom and significance levels.

Steps of Hypothesis Testing

Let us review the chi-square test for goodness of fit. We will use the same example, but this time systematically follow our standard five steps of hypothesis testing. In the process we also consider some fine points.

❶ **Restate the question as a research hypothesis and a null hypothesis about the populations.** There are two populations:

Population 1: Characters on cereal boxes like those in the study.
Population 2: Characters on cereal boxes who are equally likely to be male and female.

TIP FOR SUCCESS

Remember that Population 2 is the population that is consistent with the null hypothesis being true. Since this study focuses on the proportion of male and female characters on cereal boxes, Population 2 consists of characters on cereal boxes who are equally likely to be male or female.

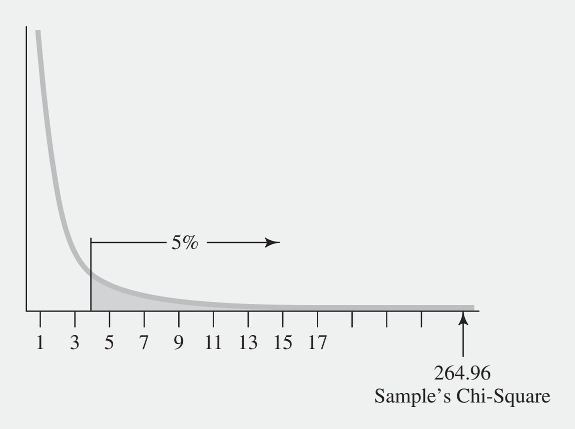


Figure 2 For the Black et al. (2009) example, the chi-square distribution ($df = 1$) showing the cutoff for rejecting the null hypothesis at the .05 level and the sample's chi-square (note that the sample's chi-square would, in reality, be located even further to the right).

TIP FOR SUCCESS

It is important not to be confused by the terminology here. The comparison distribution is the distribution to which we compare the number that summarizes the whole pattern of the result. With a *t* test, this number is the *t* score, and we use a *t* distribution. With an analysis of variance, it is the *F* ratio and we use an *F* distribution. Accordingly, with a chi-square test, our comparison distribution is a distribution of the chi-square statistic. This can be confusing because when preparing to use the chi-square distribution, we compare a distribution of observed frequencies to a distribution of expected frequencies. Yet the distribution of expected frequencies is not a comparison distribution in the sense that we use this term in Step ❷ of hypothesis testing.

The research hypothesis is that the distribution of observations over categories in the two populations is different; the null hypothesis is that they are the same.

❷ **Determine the characteristics of the comparison distribution.** The comparison distribution here is a chi-square distribution with 1 degree of freedom. (Once you know the total, there is only one category number still free to vary.)

❸ **Determine the cutoff on the comparison distribution at which the null hypothesis should be rejected.** You do this by looking up the cutoff on the chi-square table for your significance level and the study's degrees of freedom. In the present example, we used the .05 significance level, and we determined in Step ❷ that there was 1 degree of freedom. Based on the chi-square table, this gives a cutoff chi-square of 3.841.

❹ **Determine your sample's score on the comparison distribution.** Your sample's score is the chi-square figured from the sample. In other words, this is where you do all the figuring.

Ⓐ **Determine the actual, observed frequencies in each category.** These are shown in the first column of Table 1.

Ⓑ **Determine the expected frequencies in each category.** We figured these each to be 693 based on expecting an equal distribution of the 1,386 characters.

Ⓒ **In each category, take observed minus expected frequencies.** These are shown in the third column of Table 1.

Ⓓ **Square each of these differences.** These are shown in the fourth column of Table 1.

Ⓔ **Divide each squared difference by the expected frequency for its category.** These are shown in the fifth column of Table 1.

Ⓕ **Add up the results of Step Ⓔ for all the categories.** The result we figured earlier (264.96) is the chi-square statistic for the sample. It is shown in Figure 2.

❺ **Decide whether to reject the null hypothesis.** The chi-square cutoff to reject the null hypothesis (from Step ❸) is 3.841 and the chi-square of the sample (from Step ❹) is 264.96. Thus, you can reject the null hypothesis. The research hypothesis that the two populations are different is supported. That is, consistent with their hypothesis, Black et al. (2009) could conclude that the characters on cereal boxes are more likely to be male than female. Now take a look at the cereal boxes you have at home and see if there are more male than female characters on the boxes. Be sure to look at all of the characters on both the front and back of the box!

How are you doing?

1. In what situation do you use a chi-square test for goodness of fit?
2. List the steps for figuring the chi-square statistic, and explain the logic behind each step.
3. Write the formula for the chi-square statistic and define each of the symbols.
4. (a) What is a chi-square distribution? (b) What is its shape? (c) Why does it have that shape?
5. Use the steps of hypothesis testing to carry out a chi-square test for goodness of fit (using the .05 significance level) for a sample in which one category has 15 people, the other category has 35 people, and the first category is expected to have 60% of people and the second category is expected to have 40% of people.

Answers

1. You use a chi-square test for goodness of fit when you want to test whether a sample's distribution of people (or observations) across categories represents a population that is significantly different from a population with an expected distribution of people (or observations) across categories.

2. Ⓐ **Determine the actual, observed frequencies in each category.** This is the key information for the sample studied.

   Ⓑ **Determine the expected frequencies in each category.** Having these numbers makes it possible to make a direct comparison of what is expected to the observed frequencies.

   Ⓒ **In each category, take observed minus expected frequencies.** This is the direct comparison of the distribution for the sample versus the distribution representing the expected population.

   Ⓓ **Square each of these differences.** This gets rid of the direction of the difference (since the interest is only in how much difference there is).

   Ⓔ **Divide each squared difference by the expected frequency for its category.** This adjusts the degree of difference for the absolute size of the expected frequencies.

   Ⓕ **Add up the results of Step Ⓔ for all the categories.** This gives you a statistic for the overall degree of discrepancy.

3. Formula for the chi-square statistic: $\chi^2 = \Sigma\frac{(O - E)^2}{E}$.

   χ^2 is the chi-square statistic; Σ tells you to sum over all the different categories; O is the observed frequency for a category; E is the expected frequency for a category.

4. (a) A chi-square distribution: for any particular number of categories, the distribution you would expect if you figured a very large number of chi-square statistics for samples from a population in which the distribution of people over categories is the expected distribution.

   (b) It is skewed to the right.

   (c) It has this shape because a chi-square statistic can't be less than 0 (since the numerator, a squared score, has to be positive, and its denominator, an expected number of individuals, also has to be positive), but there is no limit to how large it can be.

5. ❶ **Restate the question as a research hypothesis and a null hypothesis about the populations.** There are two populations:

   Population 1: People like those in the sample.
   Population 2: People in general who have a distribution of 60% in the first category and 40% in the second category.

   The research hypothesis is that the distribution of numbers of people over categories in the population is 60% in the first category and 40% in the second category; the null hypothesis is that the distribution is not 60% in the first category and 40% in the second category.

   ❷ **Determine the characteristics of the comparison distribution.** The comparison distribution is a chi-square distribution with 1 degree of freedom (that is, $df = N_{\text{Categories}} - 1 = 2 - 1 = 1$).

❸ **Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected.** At the .05 level with *df* = 1, cutoff is 3.841.

❹ **Determine your sample's score on the comparison distribution.**

- Ⓐ **Determine the actual, observed frequencies in each category.** As given in the problem, these are 15 and 35.
- Ⓑ **Determine the expected frequencies in each category.** With 50 people total and expecting a 60% to 40% breakdown, the expected frequencies are 30 and 20.
- Ⓒ **In each category, take observed minus expected frequencies.** These come out to -15 (that is, $15 - 30 = -15$) and 15 (that is, $35 - 20 = 15$).
- Ⓓ **Square each of these differences.** Both come out to 225 (that is, $-15^2 = 225$ and $15^2 = 225$).
- Ⓔ **Divide each squared difference by the expected frequency for its category.** For the first category, this comes out to 7.5 (that is, $225/30 = 7.5$). For the second category, this comes out to 11.25 (that is, $225/20 = 11.25$).
- Ⓕ **Add up the results of Step Ⓔ for all the categories.** $7.5 + 11.25 = 18.75$.

❺ **Decide whether to reject the null hypothesis.** The sample's chi-square of 18.75 is more extreme than the cutoff of 3.841. Reject the null hypothesis; people like those in the sample are different from the expected 60% to 40% breakdown.

The Chi-Square Test for Independence

So far, we have looked at the distribution of *one nominal variable* with several categories, such as the gender breakdown of characters on cereal boxes. In fact, this kind of situation is fairly rare in research. We began with an example of this kind mainly as a stepping stone to a more common actual research situation, to which we now turn.

The most common use of chi-square is one in which there are *two nominal variables,* each with several categories. Hypothesis testing in this kind of situation is called a *chi-square test for independence.* You will learn shortly why it has this name. For example, in addition to being interested in the gender breakdown of characters on cereal boxes, Black et al. (2009) were also interested in the age distribution of the characters (in terms of whether they were children or adults). In this case, we have two nominal variables. The gender of the characters (that is, male vs. female) is the first nominal variable. The age of the characters (that is, child vs. adult) is the second nominal variable. Based on previous research, Black et al. hypothesized that male characters would be more likely to be displayed as adults than female characters. Table 3 shows the results. Notice the two nominal variables: *gender* (with two levels) and *age* (with two levels). For this part of their study, the researchers focused on 222 characters for which the student research assistants carried out detailed coding of both gender and age.

Contingency Tables

contingency table Two-dimensional chart showing frequencies in each combination of categories of two nominal variables, as in a chi-square test for independence.

Table 3 is a **contingency table**—a table in which the distributions of two nominal variables are set up so that you have the frequencies of their combinations, as well as the totals. Thus, in Table 3, the 125 in the Male/Adult combination is how many

Table 3 Contingency Table of Observed Frequencies of Gender and Age of Characters on Cereal Boxes (Data from Black et al., 2009)

		Gender		Total
		Male	Female	
Age	Child	28	30	58 (26.1%)
	Adult	125	39	164 (73.9%)
	Total	153	69	222 (100%)

of the male characters were adults. (A contingency table is similar to the tables in factorial analysis of variance where each cell had a mean of the scores of several people; but in a contingency table, the number in each cell is not a mean, but rather the *number of people or observations* that have a combination of variables.)

Table 3 is called a 2 × 2 contingency table because it has two levels of one variable crossed with two levels of the other (which dimension is named first does not matter). It is also possible to have larger contingency tables, such as a 2 × 3 table, or even 4 × 7 or 6 × 18. Smaller ones, like this 2 × 2 contingency table, are especially common.

Independence

The question in this example is whether there is any relation between the gender of the characters and whether they are children or adults. If there is no relation, the *proportion* of child and adult characters is the same among male characters and female characters. Or to put it the other way, if there is no relation, the *proportion* of male and female characters is the same for child and adult characters. However you describe it, the situation of no relation between the variables in a contingency table is called **independence.**[1]

Sample and Population

In the observed survey results in the example, the proportions of child and adult characters vary for male and female characters. For male characters, 18.3% are child characters and 81.7% are adult characters. However, among female characters, 43.5% are child characters and 56.5% are adult characters. Thus, as hypothesized by the researchers, more of the male characters are adults than are the female characters. Still, the sample is only of 222 characters. Thus, it is possible that in the larger population of cereal boxes, the age of characters is independent of the characters' being male or female. The big question is whether the lack of independence in the sample is large enough to reject the null hypothesis of independence in the population. That is, you need to do a chi-square test.

Determining Expected Frequencies

One thing that is new in a chi-square test for independence is that you now have to figure differences between observed and expected for each *combination* of categories—that is, for each **cell** of the contingency table. (When there was only one nominal

independence Situation of no relationship between two variables; term usually used regarding two nominal variables in the chi-square test for independence.

cell In chi-square, the particular combination of categories for two variables in a contingency table.

[1] Independence is usually used to talk about a lack of relation between two nominal variables. However, it may be helpful to think of independence as roughly the same as the situation of no correlation ($r = 0$).

TIP FOR SUCCESS

Always ensure that you have the same number of expected frequencies as observed frequencies. For example, with a 2 × 2 contingency table, there will be four observed frequencies and four corresponding expected frequencies. In Table 4, we have put in the expected frequency (in parentheses) next to each observed frequency.

Table 4 Contingency Table of Observed (and Expected) Frequencies of Gender and Age of Characters on Cereal Boxes (Data from Black et al., 2009)

		Gender		Total
		Male	Female	
Age	Child	28 (39.9)[a]	30 (18.0)	58 (26.1%)
	Adult	125 (113.1)	39 (51.0)	164 (73.9%)
	Total	153	69	222 (100%)

[a] Expected frequencies are in parentheses.

variable, you figured these differences just for each category of that single nominal variable.) Table 4 is the contingency table for the example of characters on cereal boxes with the expected frequency shown (in parentheses) for each cell.

The key idea to keep in mind when figuring expected frequencies in a contingency table is that "expected" is based on the two variables being independent. If they are independent, then the proportions up and down the cells of each column should be the same. In the example, overall, there are 26.1% child characters and 73.9% adult characters; thus, if character gender is independent of being a child or adult character, this 26.1–73.9% split should hold for each column (male and female). First, the 26.1–73.9% overall split should hold for the male characters. This would make an expected frequency in the male cell for child characters of 26.1% of 153 (the total number of male characters), which comes out to 39.9. (Don't worry that 39.9 isn't a whole number; even though you can't have 0.9 of a character, it makes sense because it's an expected, or theoretical, value.) The expected frequency for the male characters that are adults is 113.1 (that is, 73.9% of 153 is 113.1). The same principle holds for the column of female characters: The 69 female characters should have a 26.1–73.9% split, giving an expected frequency of 18.0 child characters (that is, 26.1% of 69 is 18.0) and 51.0 adult characters (that is, 73.9% of 69 is 51.0).

Summarizing what we have said in terms of steps,

TIP FOR SUCCESS

Be sure to check that you are selecting the correct row percentage and column total for each cell. Selecting the wrong value is a common mistake when figuring chi-square.

i **Find each row's percentage of the total.**

ii **For each cell, multiply its row's percentage by its column's total.**

Applying these steps to the top left cell (child characters who are male),

i **Find each row's percentage of the total.** The 58 characters in the child row is 26.1% of the overall total of 222 (that is, 58/222 = 26.1%).

ii **For each cell, multiply its row's percentage by its column's total.** The column total for the male characters is 153; 26.1% of 153 comes out to 39.9 (that is, .261 × 153 = 39.9).

These steps can also be stated as a formula,

A cell's expected frequency is the number in its row divided by the total number of people, multiplied by the number in its column.

$$E = \left(\frac{R}{N}\right)(C) \qquad (3)$$

In this formula, E is the expected frequency for a particular cell, R is the number of people (or observations) observed in this cell's row, and N is the total number of people (or observations) (thus, R divided by N is the proportion of the total number of people or observations that are in that row). *C* is the number of people (or observations) in this cell's column.

Applying the formula to the same top left cell,

$$E = \left(\frac{R}{N}\right)(C) = \left(\frac{58}{222}\right)(153) = (.261)(153) = 39.9$$

Looking at the entire Table 4, notice that in each column (as well as overall) the expected frequencies add up to the same totals as the observed frequencies. (This is as it should be because the expected frequencies are just a different way of dividing up the column total.) For example, in the first column (male characters), the expected frequencies of 39.9 and 113.1 add up to 153, just as the observed frequencies in that column of 28 and 125 do. Similarly, in the top row (child characters), the expected frequencies of 39.9 and 18 add up to 57.9, the same total (allowing for rounding error) as for the observed frequencies of 28 and 30.

TIP FOR SUCCESS

As a check on your arithmetic, it is a good idea to make sure that the expected and observed frequencies add up to the same row and column totals.

Figuring Chi-Square

You figure chi-square the same way as in the chi-square test for goodness of fit, except that you now figure the weighted squared difference for each *cell* and add these up. Here is how it works for our example:

$$\chi^2 = \Sigma\frac{(O-E)^2}{E} = \frac{(28-39.9)^2}{39.9} + \frac{(30-18)^2}{18} + \frac{(125-113.1)^2}{113.1} + \frac{(39-51)^2}{51}$$

$$= 3.55 + 8 + 1.25 + 2.82 = 15.62.$$

Degrees of Freedom

A contingency table with many cells may have relatively few degrees of freedom. In our example, there are four cells but only 1 degree of freedom. Recall that the degrees of freedom are the number of categories free to vary once the totals are known. With a chi-square test for independence, the number of categories is the number of cells; the totals include the row and column totals—and if you know the row and column totals, you have a lot of information.

Consider our example of the gender and age of characters on cereal boxes. Suppose you know the first cell frequency across the top, for example, and all the row and column totals. You could then figure all the other cell frequencies just by subtraction. Table 5 shows the contingency table for this example with just the row and column totals and this one cell frequency. Let's start with the Child/Female cell. There are 58

Table 5 Contingency Table Showing Marginal and One Cell's Observed Frequencies to Illustrate Figuring Degrees of Freedom (Data from Black et al., 2009)

		Gender		Total
		Male	Female	
Age	Child	28	—	58 (26.1%)
	Adult	—	—	164 (73.9%)
	Total	153	69	222 (100%)

child characters in total, and the other child cell has 28 in it. Thus, only 30 remain for the Child/Female cell. Now consider the two adult character cells. You know the frequencies for all child character cells and the column totals for male and female characters. Thus, each cell frequency for the adult characters is its column's total minus the child characters in that column. For example, there are 153 male characters and 28 of those are child characters. Thus, the remaining 125 must be adult characters.

What you can see in all this is that with knowledge of only one of the cells, you could figure out the frequencies in each of the other cells. Thus, although there are four cells, there is only 1 degree of freedom—only two cells whose frequencies are really free to vary once we have all the row and column totals.

However, rather than having to think all this out each time, there is a shortcut. In a chi-square test for independence, the degrees of freedom are the number of columns minus 1 multiplied by the number of rows minus 1. Put as a formula,

The degrees of freedom for the chi-square test for independence are the number of columns minus 1 multiplied by the number of rows minus 1.

$$df = (N_{\text{Columns}} - 1)\,(N_{\text{Rows}} - 1) \tag{4}$$

N_{Columns} is the number of columns and N_{Rows} is the number of rows.

Using this formula for our survey example,

$$df = (N_{\text{Columns}} - 1)\,(N_{\text{Rows}} - 1) = (3 - 1)\,(2 - 1) = (2)\,(1) = 2.$$

With 1 degree of freedom, Table 2 shows that the chi-square you need for significance at the .05 level is 3.841. The chi-square of 15.62 for our example is larger than this cutoff. Thus, you can reject the null hypothesis that the two variables are independent in the population.

Steps of Hypothesis Testing

Now let's go through the survey example again, this time following the steps of hypothesis testing.

❶ **Restate the question as a research hypothesis and a null hypothesis about the populations.** There are two populations:

Population 1: Characters on cereal boxes like those in the study.
Population 2: Characters on cereal boxes for which the age distribution of the characters is independent of the gender of the characters.

The null hypothesis is that the two populations are the same—that, in general, the proportion of characters that are children and adults is the same for male characters and female characters. The research hypothesis is that the two populations are different, that for characters on cereal boxes, the proportion of characters that are children and adults is different for male and female characters.

Put another way, the null hypothesis is that the two variables are independent (they are unrelated to each other). The research hypothesis is that they are not independent (that they are related to each other).

❷ **Determine the characteristics of the comparison distribution.** The comparison distribution is a chi-square distribution with 1 degree of freedom (the number of columns minus 1 multiplied by the number of rows minus 1).

❸ **Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected.** You use the same table as for any chi-square test. In the example, setting a .05 significance level with 1 degree of freedom, you need a chi-square of 3.841.

❹ **Determine your sample's score on the comparison distribution.**

Ⓐ **Determine the actual, observed frequencies in each cell.** These are the results of the survey, as given in Table 3.

Ⓑ **Determine the expected frequencies in each cell.** These are shown in Table 4. For example, for the bottom right cell (Adult/Female cell):

ⓘ **Find each row's percentage of the total.** The 164 characters in the adult row are 73.9% of the overall total of 222 (that is, $164/222 = 73.9\%$).

ⓘⓘ **For each cell, multiply its row's percentage by its column's total.** The column total for female characters is 69; 73.9% of 69 comes out to 51.0 (that is, $.739 \times 69 = 51.0$).

Ⓒ **In each cell, take observed minus expected frequencies.** For example, for the Adult/Female cell, this comes out to -12 (that is, $39 - 51 = -12$).

Ⓓ **Square each of these differences.** For example, for the Adult/Female cell, this comes out to 144 (that is, $-12^2 = 144$).

Ⓔ **Divide each squared difference by the expected frequency for its cell.** For example, for the Adult/Female cell, this comes out to 5.14 (that is, $144/51 = 2.82$).

Ⓕ **Add up the results of Step Ⓔ for all the cells.** As we saw, this came out to 15.62.

❺ **Decide whether to reject the null hypothesis.** The chi-square needed to reject the null hypothesis is 3.841 and the chi-square for our sample is 15.62 (see Figure 3). Thus, you can reject the null hypothesis. The research hypothesis that the two variables are not independent in the population is supported. That is, the proportion of characters that are children or adults is different for male and female characters.

A Second Example

Riehl (1994) studied the college experience of first-year students who were the first generation in their family to attend college. These students were compared to other students who were not the first generation in their family to go to college. (All students in the study were from Indiana University.) One of the variables Riehl measured was whether students dropped out during their first semester.

Table 6 shows the results along with the expected frequencies (shown in parentheses) based on these percentages. Below the contingency table is the figuring for the chi-square test for independence.

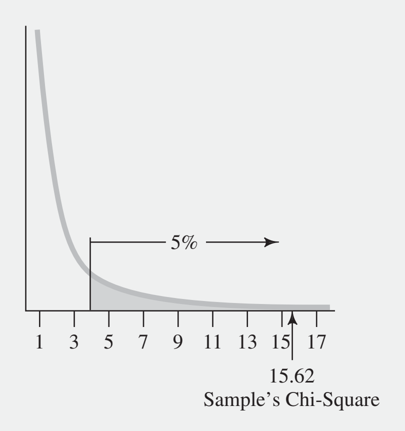


Figure 3 For the Black et al. (2009) example of the age and gender of characters on cereal boxes, the chi-square distribution ($df = 1$) showing the cutoff for rejecting the null hypothesis at the .05 level and the sample's chi-square.

Table 6 Results and Figuring of the Chi-Square Test for Independence Comparing Whether First-Generation College Students Differ from Others in First-Semester Dropouts

	Generation to Go to College		**Total**
	First Ⓐ	Other Ⓑ	
Dropped Out	73 (57.7)	89 (103.9)	162 (7.9%)
Did Not Drop Out	657 (672.3)	1,226 (1,211.1)	1,883 (92.1%)
Total	730	1,315	2,045

$df = (N_{\text{Columns}} - 1)(N_{\text{Rows}} - 1) = (2 - 1)(2 - 1) = (1)(1) = 1.$ ❷

Chi-square needed, $df = 1$, .01 level: 6.635. ❸

$$\chi^2 = \Sigma \frac{(O - E)^2}{E} = \frac{(73 - 57.7)^2}{57.7} + \frac{(89 - 103.9)^2}{103.9} + \frac{(657 - 672.3)^2}{672.3} + \frac{(1{,}226 - 1{,}211.1)^2}{1{,}211.1}$$

Ⓒ
$$= \frac{15.3^2}{57.7} + \frac{-14.9^2}{103.9} + \frac{-15.3^2}{672.3} + \frac{14.9^2}{1{,}211.1}$$

Ⓓ
$$= \frac{234.1}{57.7} + \frac{222}{103.9} + \frac{234.1}{672.3} + \frac{222}{1{,}211.1}$$

Ⓔ
$$= 4.06 + 2.14 + .35 + .18$$

$$= 6.73.$$ Ⓕ

Decision: Reject the null hypothesis. ❺

Notes. 1. With a 2 × 2 analysis, the differences and squared differences (numerators) are the same for all four cells. In this example, the cells are a little different due to rounding error. 2. Data from Riehl (1994). The exact chi-square (6.73) is slightly different from that reported in the article (7.2), due to rounding error.

❶ **Restate the question as a null hypothesis and a research hypothesis about the populations.** There are two populations:

Population 1: Students like those surveyed.
Population 2: Students whose dropping out or staying in college their first semester is independent of whether they are the first generation in their family to go to college.

The null hypothesis is that the two populations are the same—that, in general, whether students drop out of college is independent of whether they are the first generation in their family to go to college. The research hypothesis is that the populations are not the same. In other words, the research hypothesis is that students like those surveyed are unlike the hypothetical population in which dropping out is unrelated to whether you are first generation.

❷ **Determine the characteristics of the comparison distribution.** This is a chi-square distribution with 1 degree of freedom.

❸ **Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected.** Using the .01 level and 1 degree of freedom, you need a chi-square for significance of 6.635. This is shown in Figure 4.

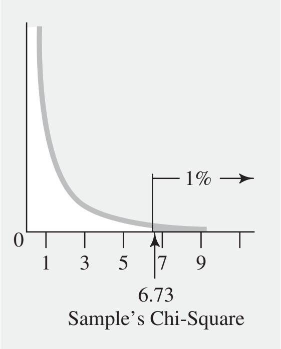


Figure 4 For the example from Riehl (1994), chi-square distribution ($df = 1$) showing the cutoff for rejecting the null hypothesis at the .01 level and the sample's chi-square.

❹ **Determine your sample's score on the comparison distribution.**

Ⓐ **Determine the actual, observed frequencies in each cell.** These are the results of the survey, as given in Table 6.

Ⓑ **Determine the expected frequencies in each cell.** These are shown in parentheses in Table 6. For example, for the top left cell (First Generation/Dropped Out cell),

ⓘ **Find each row's percentage of the total.** The Dropped Out row's 162 is 7.9% of the overall total of 2,045 (that is, 162/2,045 = 7.9%).

ⓘⓘ **For each cell, multiply its row's percentage by its column's total.** The column total for the First Generation students is 730; 7.9% of 730 comes out to 57.7 (that is, .079 × 730 = 57.7).

Ⓒ **In each cell, take observed minus expected frequencies.** These are shown in Table 6. For example, for the First Generation/Dropped Out cell, this comes out to 15.3 (that is, 73 − 57.7 = 15.3).

Ⓓ **Square each of these differences.** These are also shown in Table 6. For example, for the First Generation/Dropped Out cell, this comes out to 234.1 (that is, $15.3^2 = 234.1$).

Ⓔ **Divide each squared difference by the expected frequency for its cell.** Once again, these are shown in Table 6. For example, for the First Generation/Dropped Out cell, this comes out to 4.06 (that is, 234.1/57.7 = 4.06).

Ⓕ **Add up the results of Step Ⓔ for all the cells.** As shown in Table 6, this comes out to 6.73. Its location on the chi-square distribution is shown in Figure 4.

❺ **Decide whether to reject the null hypothesis.** Your chi square of 6.73 is larger than the cutoff of 6.635 (see Figure 4). Thus, you can reject the null hypothesis. That is, judging from a sample of 2,045 Indiana University students, first-generation students are somewhat more likely to drop out during their first semester than are other students. (Remember, of course, that there could be many reasons for this result.) The results of Riehl's (1994) study have been replicated in many other universities. For example, a study of this issue among college students across the United States also found that first-generation students were more likely to drop out of college than other students (Chen, 2005).

How are you doing?

1. (a) In what situation do you use a chi-square test for independence? (b) How is this different from the situation in which you would use a chi-square test for goodness of fit?
2. (a) List the steps for figuring the expected frequencies in a contingency table. (b) Write the formula for expected frequencies in a contingency table and define each of its symbols.
3. (a) Write the formula for figuring degrees of freedom in a chi-square test for independence and define each of its symbols. (b) Explain the logic behind this formula.
4. Use the steps of hypothesis testing to carry out a chi-square test for independence for the following observed scores (using the .01 significance level). The scores are from a study in which 200 university staff members completed a survey about the kind of transportation they use to get to work and whether they are "morning people" (prefer to go to bed early and awaken early) or "night people" (go to bed late and awaken late).

		Transportation			
		Bus	Carpool	Own Car	Total
Sleep Tendency	Morning	60	30	30	120
	Night	20	20	40	80
	Total	80	50	70	200

Answers

1. (a) You use a chi-square test for independence when you have the number of people in each of the various combinations of levels of two nominal variables, and you want to test whether the difference from independence in the sample is sufficiently great to reject the null hypothesis of independence in the population. (b) The focus is on the independence of *two nominal variables,* whereas in a chi-square test for goodness of fit the focus is on the distribution of people over categories of a *single nominal variable.*

2. (a) ⅰ Find each row's percentage of the total.
   ⅱ For each cell, multiply its row's percentage by its column's total.

   (b) $E = \left(\frac{R}{N}\right)(C)$

   E is the expected frequency for a particular cell, R is the number of people observed in this cell's row, N is the total number of people, and C is the number of people observed in this cell's column.

3. (a) $df = (N_{\text{Columns}} - 1)\ (N_{\text{Rows}} - 1)$.

   df are the degrees of freedom, N_{Columns} is the number of columns, and N_{Rows} is the number of rows.

   (b) df are the number of cell totals free to vary, given you know the column and row totals. If you know the totals in all the columns but one, you can figure the total in the cells in the remaining column by subtraction. Similarly, if you know the total in all the rows but one, you can figure the total in the cells in the remaining row by subtraction.

4. ❶ **Restate the question as a null hypothesis and a research hypothesis about the populations.** There are two populations:

   Population 1: People like those studied.
   Population 2: People for whom being a morning or night person is independent of the kind of transportation they use to commute to work.

   The null hypothesis is that the two populations are the same (that, in general, the proportions using different types of transportation are the same for morning and night people); the research hypothesis is that the two populations are not the same (that among people in general the proportions using different types of transportation are different for morning and night people).

   ❷ **Determine the characteristics of the comparison distribution.** This is a chi-square distribution with 2 degrees of freedom. That is, $df = (N_{\text{Columns}} - 1)(N_{\text{Rows}} - 1) = (3 - 1)(2 - 1) = 2$.

   ❸ **Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected.** From Table 4 of the appendix "Tables," for the .01 significance level and 2 degrees of freedom, the needed chi-square is 9.211.

   ❹ **Determine your sample's score on the comparison distribution.**

   Ⓐ **Determine the actual, observed frequencies in each cell.** These are shown in the contingency table for the problem.

   Ⓑ **Determine the expected frequencies in each cell.** For example, for the bottom right cell (Night Person/Own Car cell),

   ⅰ **Find each row's percentage of the total.** The 80 people in the night person's row are 40% of the overall total of 200 (that is, 80/200 = 40%).

ⓘⓘ **For each cell, multiply its row's percentage by its column's total.** The column total for those with their own car is 70; 40% of 70 comes out to 28 (that is, .40 × 70 = 28).

Ⓒ **In each cell, take observed minus expected frequencies.** For example, for the Night Person/Own Car cell, this comes out to 12 (that is 40 − 28 = 12).

Ⓓ **Square each of these differences.** For example, for the Night Person/Own Car cell, this comes out to 144 (that is, $12^2 = 144$).

Ⓔ **Divide each squared difference by the expected frequency for its cell.** For example, for the Night Person/Own Car cell, this comes out to 5.14 (that is, 144/28 = 5.14).

Ⓕ **Add up the results of Step Ⓔ for all the cells.** 3 + 0 + 3.43 + 4.50 + 0 + 5.14 = 16.07.

❺ **Decide whether to reject the null hypothesis.** The sample's chi square of 16.07 is larger than the cutoff of 9.211. Thus, you can reject the null hypothesis. The research hypothesis that the two variables are not independent in the population is supported. That is, the proportions of type of transportation used to commute to work are different for morning and night people.

Assumptions for the Chi-Square Tests

The chi-square tests for goodness of fit and for independence do not require the usual assumptions of normal population variances and such. There is, however, one key assumption: Each score must not have any special relation to any other score. This means that you can't use these chi-square tests if the scores are based on the same people being tested more than once. Consider a study in which 20 people are tested to see if the distribution of their preferred brand of breakfast cereal changed from before to after a recent nutritional campaign. The results of this study could not be tested with the usual chi-square because the distributions of cereal choice before and after are from the same people.

Effect Size and Power for Chi-Square Tests for Independence

Effect Size

In chi-square tests for independence, you can use your sample's chi-square to figure a number that shows the degree of association of the two nominal variables.

With a 2 × 2 contingency table, the measure of association is called the **phi coefficient (ϕ).** Here is the formula,

$$\phi = \sqrt{\frac{\chi^2}{N}} \qquad (5)$$

The phi coefficient (effect size for a chi-square test for independence for a 2 × 2 contingency table) is the square root of the result of dividing the sample's chi-square by the total number of people in the sample.

The phi coefficient has a minimum of 0 and a maximum of 1, and can be considered similar to a correlation coefficient[2] Cohen's (1988) conventions for

phi coefficient (ϕ) Measure of association between two dichotomous nominal variables; square root of division of chi-square statistic by *N;* equivalent to correlation of the two variables if they were each given numerical values (for example, of 1 and 0 for the two categories); effect-size measure for a chi-square test of independence with a 2 × 2 contingency table.

[2] Phi is actually identical to the correlation coefficient. Suppose you were to take the two variables in a 2 × 2 contingency table and arbitrarily make one of the values of each equal to 1 and the other equal to 2 (you could use any two different numbers). And suppose you then figured a correlation coefficient between the two variables. The result would be exactly the same as the phi coefficient. (Whether it was a positive or negative correlation, however, would depend on which categories in each variable got the higher number.)

the phi coefficient are that .10 is a small effect size, .30 is a medium effect size, and .50 is a large effect size (the same as for a correlation coefficient).

For example, in the Riehl (1994) study of first-generation college students, the chi-square we calculated was 6.73 and there were 2,045 people in the study. Applying the formula for the phi coefficient,

$$\phi = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{6.73}{2{,}045}} = \sqrt{.00329} = .06$$

This is a very small effect size. The fact that the chi-square of 6.73 was significant in this study tells you that the greater likelihood of first-generation students dropping out that you saw in the sample is probably not due to the particular people that were randomly recruited to be in this sample. You can thus have some confidence that there is a pattern of this kind in the population. But the small phi coefficient tells you that this true population tendency may not be a very important factor in practice.

You only use the phi when you have a 2 × 2 situation. **Cramer's phi** is an extension of the ordinary phi coefficient that you can use for contingency tables larger than 2 × 2. (Cramer's phi is also known as *Cramer's V* and is sometimes written as ϕ_C or V_C.) You figure Cramer's phi the same way as the ordinary phi coefficient, except that instead of dividing by *N,* you divide by *N* times the degrees of freedom of the smaller side of the table. Stated as a formula,

Cramer's phi coefficient (effect size for a chi-square test for independence) is the square root of the result of dividing the sample's chi-square by the product of the total number of people in the sample times the degrees of freedom for the smaller side of the table.

$$\text{Cramer's } \phi = \sqrt{\frac{\chi^2}{(N)(df_{\text{Smaller}})}} \qquad (6)$$

In this formula, df_{Smaller} is the degrees of freedom for the smaller side of the contingency table.

For example, consider the result of a chi-square test for independence in which the sample's chi-square was 16.07, the total number of people surveyed was 200, and the study used a 3 × 2 contingency table. The degrees of freedom for the smaller side of the table was 1. Cramer's phi is the square root of what you get when you divide 16.07 by 200 multiplied by 1. This comes out to .28. In terms of the formula,

$$\text{Cramer's } \phi = \sqrt{\frac{\chi^2}{(N)(df_{\text{Smaller}})}} = \sqrt{\frac{16.07}{(200)(1)}} = \sqrt{.08} = .28$$

Cramer's phi Effect-size measure for a chi-square test for independence used with a contingency table that is larger than 2 × 2; square root of result of dividing the chi-square statistic by the product of the number of participants times the degrees of freedom of the smaller side of the contingency table; also known as *Cramer's V* and sometimes written as ϕ_c or V_c.

Cohen's (1988) conventions for effect size for Cramer's phi depend on the degrees of freedom for the smaller side of the table. Table 7 shows Cohen's effect size conventions for Cramer's phi for tables in which the smallest side of the table is 2, 3,

Table 7 Cohen's Conventions for Cramer's Phi

Smallest Side of Contingency Table	Effect Size		
	Small	**Medium**	**Large**
2 ($df_{\text{Smaller}} = 1$)	.10	.30	.50
3 ($df_{\text{Smaller}} = 2$)	.07	.21	.35
4 ($df_{\text{Smaller}} = 3$)	.06	.17	.29

and 4. Notice that when the smallest side of the table is 2, the degrees of freedom is 1. Thus, the effect sizes in the table for this situation are the same as for the ordinary phi coefficient. Based on the table, a Cramer's phi of .28 represents an approximately medium effect size (.28).

Power

Table 8 shows the approximate power at the .05 significance level for small, medium, and large effect sizes and total sample sizes of 25, 50, 100, and 200. Power is given for tables with 1, 2, 3, and 4 degrees of freedom.[3]

Consider the power of a planned 2 × 4 study $(df = 3)$ of 50 people with an expected medium effect size (Cramer's $\phi = .30$). The researchers will use the .05 level. From Table 8 you can find that this study would have a power of .40. That is, if the research hypothesis is true, and there is a true medium effect size, there is about a 40% chance that the study will come out significant. Notice from this table two things about power for a chi-square test for independence. First, like all other hypothesis-testing situations, the more participants there are in the study, the more power there will be. Second, the more degrees of freedom there are (the more different categories are crossed with each other), the less power there is. Thus, for maximum power, you want as many participants as possible with as simple a contingency table (that is, as few categories in each direction) as possible.

Table 8 Approximate Power for the Chi-Square Test for Independence for Testing Hypotheses at the .05 Significance Level

			Effect Size	
Total *df*	**Total *N***	**Small**	**Medium**	**Large**
1	25	.08	.32	.70
	50	.11	.56	.94
	100	.17	.85	*
	200	.29	.99	*
2	25	.07	.25	.60
	50	.09	.46	.90
	100	.13	.77	*
	200	.23	.97	*
3	25	.07	.21	.54
	50	.08	.40	.86
	100	.12	.71	.99
	200	.19	.96	*
4	25	.06	.19	.50
	50	.08	.36	.82
	100	.11	.66	.99
	200	.17	.94	*

*Nearly 1.

[3] Cohen (1988, pp. 228–248) gives more detailed tables. However, Cohen's tables are based on an effect size called *w,* which is equivalent to phi but not to Cramer's phi. He provides a helpful conversion table of Cramer's phi to *w* on page 222.

Table 9 Approximate Total Number of Participants Needed for 80% Power for the Chi-Square Test for Independence for Testing Hypotheses at the .05 Significance Level

	Effect Size		
Total *df*	**Small**	**Medium**	**Large**
1	785	87	26
2	964	107	39
3	1,090	121	44
4	1,194	133	48

Needed Sample Size

Table 9 gives the approximate total number of participants needed for 80% power with small, medium, and large effect sizes at the .05 significance level for chi-square tests for independence of 1, 2, 3, and 4 degrees of freedom.[4] Suppose you are planning a study with a 3 × 3 $(df = 4)$ contingency table. You expect a large effect size and will use the .05 significance level. According to the table, you would only need 48 participants. Again, the same principle holds that we emphasized earlier regarding the degrees of freedom when figuring power. In this case, it means that the more degrees of freedom there are (that is, the more categories in each variable being crossed), the more participants you need for the same amount of power.

How are you doing?

1. What are the assumptions for chi-square tests?
2. (a) What is the measure of effect size for a 2 × 2 chi-square test for independence? (b) Write the formula for this measure of effect size and define each of the symbols. (c) What are Cohen's conventions for small, medium, and large effect sizes? (d) Figure the effect size for a 2 × 2 chi-square test for independence in which there are a total of 100 participants and the chi-square is 12.
3. (a) What is the measure of effect size for a chi-square test for independence for a contingency table that is larger than 2 × 2? (b) Write the formula for this measure of effect size and define each of the symbols. (c) What is Cohen's convention for a small effect size for a 4 × 6 contingency table? (d) Figure the effect size for a 4 × 6 chi-square test for independence in which there are a total of 200 participants and the chi-square is 20.
4. What is the power of a planned 3 × 3 chi-square with 50 participants total and a predicted medium effect size?
5. What are two factors that affect the power of a study using a chi-square test of independence?
6. About how many participants do you need for 80% power in a planned 2 × 2 study in which you predict a medium effect size and will be using the .05 significance level?

[4] More detailed tables are provided in Cohen (1988, pp. 253–267). When using these tables, see footnote 3. Also, Dunlap and Myers (1997) have shown that with a 2 × 2 table, the approximate number of participants needed for 80–90% power is $8/\phi^2$.

Answers

1. The only major assumption for chi-square tests is that the numbers in each cell or category are from separate persons.
2. (a) The measure of effect size for a 2 × 2 chi-square test for independence is the phi coefficient.

   (b) The formula for the measure of effect size is $\phi = \sqrt{\frac{\chi^2}{N}}$. ϕ is the phi coefficient (effect size for a chi-square test for independence with a 2 × 2 contingency table); χ^2 is the sample's chi-square; and N is the total number of participants in the study.

   (c) Cohen's conventions: .10 is a small effect size, .30 is a medium effect size, and .50 is a large effect size.

   (d) Effect size: $\phi = \sqrt{(12/100)} = .35$.
3. (a) Cramer's phi.

   (b) Formula: Cramer's $\phi = \sqrt{\frac{\chi^2}{(N)(df_{\text{Smaller}})}}$. Cramer's ϕ is Cramer's phi coefficient (effect size for a chi-square test for independence); χ^2 is the sample's chi-square; N is the total number of participants in the study; and df_{Smaller} is the degrees of freedom for the smaller side of the contingency table.

   (c) Cohen's convention: .06.

   (d) $\sqrt{20/[(200)(3)]} = .18$.
4. Power: .36.
5. Two factors that affect the power of a study using a chi-square test of independence: number of participants and degrees of freedom. (Another factor we did not discuss here, but which also affects power for chi-square as it does for any significance test, is significance level chosen.)
6. Number of participants needed: 87.

Strategies for Hypothesis Testing When Population Distributions Are Not Normal

This second main part of the chapter examines some strategies researchers use when the variables are quantitative, but the assumption of a normal population distribution is clearly violated. (This assumption of normality is part of most ordinary hypothesis-testing procedures, such as the *t* test and the analysis of variance.) First, we briefly review the role of assumptions in the standard hypothesis-testing procedures. Then we examine two approaches researchers use when the assumptions have not been met: data transformations and rank-order tests.

Assumptions in the Standard Hypothesis-Testing Procedures

You have to meet certain conditions (the assumptions) to get accurate results with a *t* test or an analysis of variance. In these hypothesis-testing procedures, you treat the scores from a study as if they came from some larger, though unknown, populations. One assumption you have to make is that the populations involved follow a normal curve. The other main assumption you have to make is that the populations have equal variances.

That you get fairly accurate results when a study suggests that the populations even very roughly meet the assumptions of following a normal curve and having equal variances. Our concern here, however, is about the situation where there is *strong reason to believe* that the populations are nowhere near normal, or nowhere near having equal variances. In such situations, if you use the ordinary *t* test or analysis of variance, you can get quite incorrect results. For example, you could do all the figuring correctly and decide to reject the null hypothesis based on your results. Yet, if your populations do not meet the assumptions, this result could be wrong—wrong in the sense that instead of there actually being only a 5% chance of getting your results if the null hypothesis is true, in fact there might be a 15% or 20% chance! (It could also be 1% or 2%. The problem is that the usual cutoff can be a long way from accurate and you don't even know in which direction.)

Remember: Assumptions are about *populations* and not about *samples*. It is quite possible for a sample not to follow a normal curve even though it comes from a population that does follow a normal curve. Figure 5 shows histograms for several samples, each taken randomly from a population that follows a normal curve.

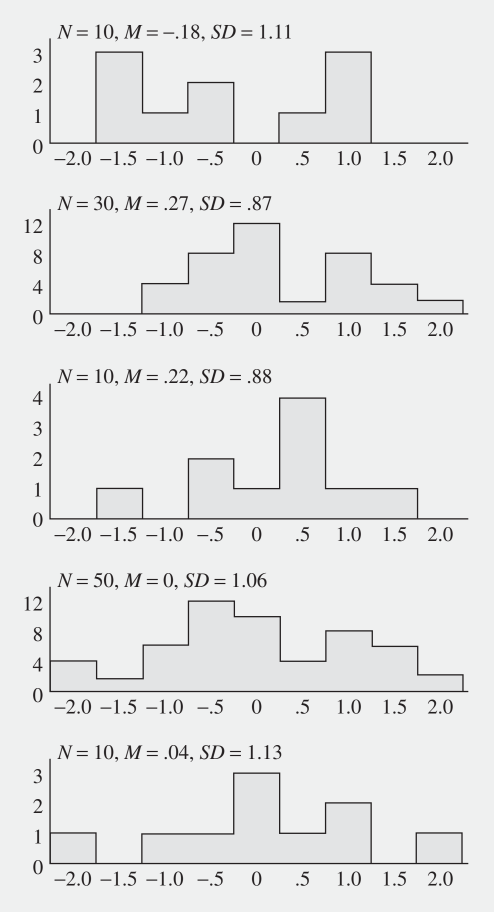


Figure 5 Histograms for several random samples, each taken from a normal population with a mean of 0 and a standard deviation of 1.

(Notice that the smaller the sample, the harder it is to see that it came from a normal population.) Of course, it is quite possible for non-normal populations to produce any of these samples as well. Unfortunately, the sample is usually all you have when doing a study. One thing researchers do is to make a histogram for the sample; if it is not drastically different from normal, the researchers assume that the population it came from is normal or at least close enough. When considering normality, most behavioral and social science researchers consider a distribution innocent (that is, normal) until proven guilty.

One common situation where you might doubt the assumption that the population follows a normal curve is when there is a *ceiling* or *floor effect*. Another common situation that raises such doubts is when the sample has outliers, extreme scores at one or both ends of the sample distribution. Figure 6 shows some examples of distributions with outliers. Outliers are a big problem in the statistical methods we ordinarily use. This is because these methods ultimately rely on squared deviations from the mean. Because it is so far from the mean, an outlier has a huge influence when you square its deviation from the mean. What this means is that a single outlier, if it is extreme enough, can drastically distort the results of a study. An outlier can cause a statistical test to give a significant result even when all the other scores would not. In other cases, an outlier can make a result not significant that would be significant without the outlier.

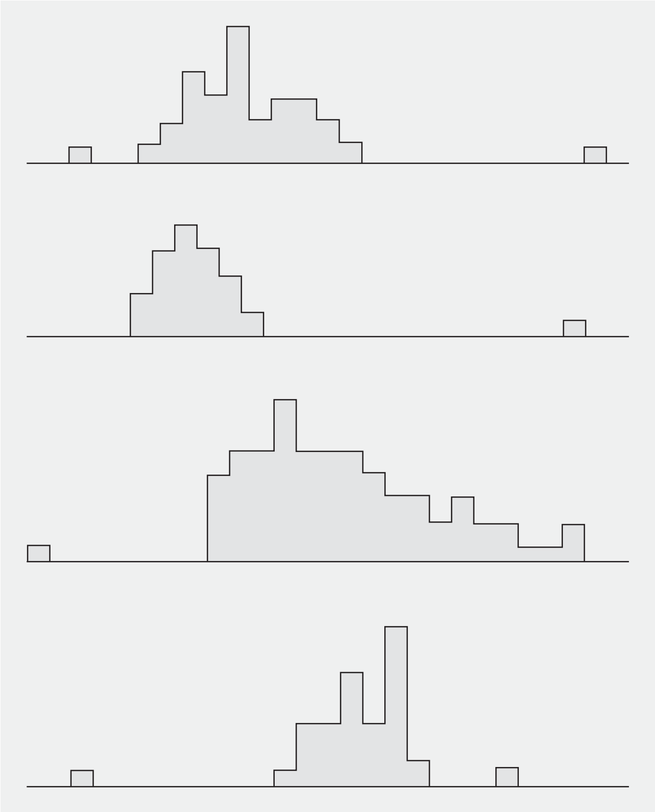


Figure 6 Distributions with outliers at one or both ends.

How are you doing?

1. What are the two main assumptions for *t* tests and the analysis of variance?
2. (a) How do you check to see if you have met the assumptions? (b) Why is this problematic?
3. (a) What is an outlier? (b) Why are outliers likely to have an especially big distorting effect in most statistical procedures?

Answers

1. The two main assumptions are that the populations are normally distributed and have equal variances.
2. (a) You look at the distributions of the samples. (b) The samples, especially if they are small, can have quite different shapes and variances from the populations.
3. (a) An outlier is an extreme score. (b) Outliers are likely to have an especially big distorting effect in most statistical procedures because most procedures are based on squared deviations from the mean. Thus, the extremeness of an outlier is greatly multiplied when its deviation from the mean is squared.

Data Transformations

One widely used procedure when the scores in the sample do not appear to come from a normal population is to change the scores! This is not done by fudging, although at first it may sound that way, until we explain. The method is that the researcher applies some mathematical procedure to each score, such as taking its square root. The idea is to make a non-normal distribution closer to normal. (Sometimes this can also make the variances of the groups more similar.) This is called a **data transformation.** Once you have made a data transformation that makes the scores in the sample appear to meet the normality assumption (and if the other assumptions are met), you can then go ahead with a usual *t* test or analysis of variance.

Data transformation has an important advantage over other procedures of coping with non-normal populations: once you have made a data transformation, you can use familiar and sophisticated hypothesis-testing procedures.

Consider an example. Measures of reaction time, such as how long it takes a research participant to press a particular key when a light flashes, are usually highly skewed to the right (positively skewed, with a long tail to the right). There are many short (quick) responses, but usually a few quite long (slow) ones. It is unlikely that the reaction times shown in Figure 7 come from a population that follows a normal curve. The population of reaction-time scores itself is probably skewed.

However, suppose you take the square root of each reaction time. Most reaction times are affected only a little. A reaction time of 1 second stays 1; a reaction time of 1.5 seconds reduces to 1.22. However, very long reaction times, the ones that create the long tail to the right, are much reduced. For example, a reaction time of 9 seconds is reduced to 3, and a reaction time of 16 seconds (for the person who was really distracted and forgot about the task) reduces to 4. (Of course, if the reaction time is as long as 16 seconds when most are around 3 or 4, we might also consider that an outlier!) Figure 8 shows the result of taking the square root of each score in the skewed distribution shown in Figure 7. After a **square-root transformation,** this distribution of scores seems much more likely to have come from a population with a normal distribution (of transformed scores).

data transformation Mathematical procedure (such as taking the square root) used on each score in a sample, usually done to make the sample distribution closer to normal.

square-root transformation Data transformation using the square root of each score.

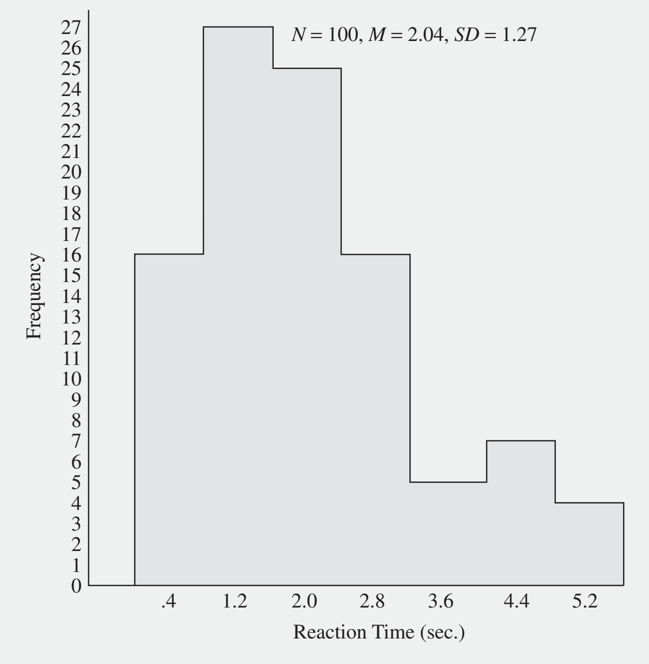


Figure 7 Skewed distribution of reaction times (fictional data).

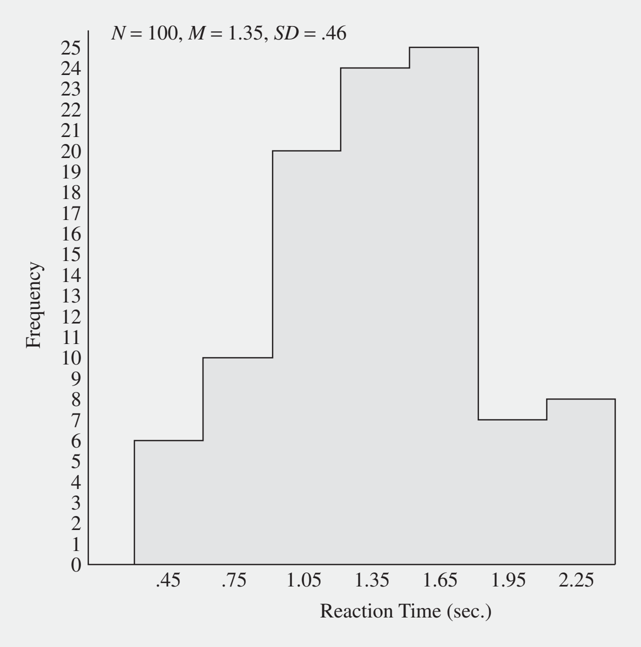


Figure 8 Distribution of scores from Figure 7 after square-root transformation.

Legitimacy of Data Transformations

Do you feel that this is somehow cheating? It would be if you did this knowingly in some way to make the result more favorable to your predictions. However, in actual research practice, the first step after the data are collected and recorded (and checked for accuracy) is to see if the data suggest that the populations meet assumptions. If the scores in your sample suggest that the populations do not meet assumptions, you do data transformations. Hypothesis testing is done only after this checking and any transformations.

Remember that you must do any transformation for *all* the scores on that variable, not just those in a particular group. Most important, no matter what transformation procedure you use, the order of the scores always stays the same. A person with an actual original score that is between the actual original scores of two other participating people will still have a transformed score between those same two people's transformed scores. (For example, consider a study in which four people have scores of 1, 2, 3, and 4; after a square root transformation, the same people have scores of 1, 1.44, 1.67, and 2. As you can see, the highest person is still highest, the second highest is still second highest, and so on.)

The procedure may seem somehow to distort reality to fit the statistics. In some cases, this is a legitimate concern. Suppose you are looking at the difference in income between two groups of Americans. You probably do not care about how much the two groups differ in the square root of their income. What you care about is the difference in actual dollars.

On the other hand, consider a survey question in which people rate their agreement with the statement "I am satisfied with local law enforcement" on a 7-point rating from 1, strongly disagree, to 7, strongly agree. Higher scores on this scale certainly mean more agreement; lower scores, less agreement. However, each scale-point increase does not necessarily mean the same amount of increase in an individual's agreement. It is just as likely that the square root of each scale point's increase is directly related to the person's underlying degree of agreement. In many research situations, there may be no strong reason to think that the transformed version is any less accurate a reflection of the reality than the original version. And the transformed version may meet the normality assumption

Kinds of Data Transformations

There are several types of data transformations. We already have shown a square-root transformation: instead of using each score, you use the square root of each score. We gave an example in Figures 7 and 8. The general effect is shown in Figure 9. As you can see, a distribution that is skewed to the right (positively skewed) becomes less skewed to the right after square-root transformation. To put it numerically, moderate numbers become only slightly lower and high numbers become much lower. The

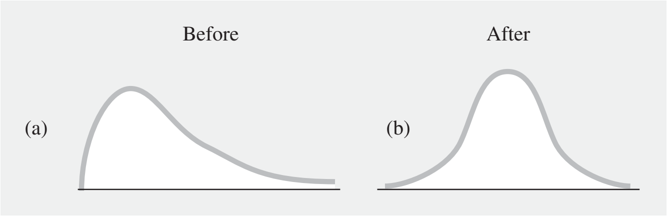


Figure 9 Distribution skewed to the right before (a) and after (b) taking the square root of each score.

result is that the right side is pulled in toward the middle. (If the distribution is skewed the other way, you may want to *reflect* all the scores—that is, subtract them all from some high number so that they are now all reversed. Then, using the square root will have the correct effect. However, you then have to remember when looking at the final results that you have reversed the direction of scoring.)

There are many other kinds of transformations you will see in behavioral and social science research articles. The square root transformation is actually quite common. Another common transformation is called a *log transformation,* in which, instead of the square root, the researcher takes the logarithm of each score. A log transformation has the same type of effect as the square root transformation, but the effect is stronger. Thus, a log transformation is better for distributions that are very strongly skewed to the right. Some other transformations you might see are *inverse transformations* and *arcsine transformations.* We will not go into examples of all these kinds of transformations here. Just learning the square-root transformation will help you get the idea. The main thing to remember about other kinds of transformations is that they all use this same principle of taking each score and applying some arithmetic to it to make the set of scores come out more like a normal curve. And as we noted before, whatever transformation you use, a score that is between two other scores always stays between those two other scores.

An Example of a Data Transformation

Consider a fictional study in which the researchers compare the number of books read in the past year by four children who score high on a test of being highly sensitive to four children who score low on the test of being highly sensitive. (The general idea of being a highly sensitive person is described in Aron, 1996, 2002; Aron & Aron, 1997.) Based on theory, the researcher predicts that highly sensitive children will read more books. Table 10 shows the results.

Table 10 Results of a Study Comparing Highly Sensitive and Not Highly Sensitive Children on the Number of Books Read in the Past Year (Fictional Data)

	Highly Sensitive	
	No	**Yes**
	0	17
	3	36
	10	45
	22	75
Σ:	35	173
$M =$	8.75	43.25
$S^2 =$	95.58	584.25

Ordinarily, in a study like this, involving a comparison of two independent groups, you would use a *t* test for independent means. Yet the *t* test for independent means is like all of the procedures you have learned for hypothesis testing (except chi-square): it requires that the parent populations of scores for each group be normally distributed. In this study, however, the distribution of the sample is strongly skewed to the right; the scores tend to bunch up at the left, leaving a long tail to the right. It thus seems likely that the population of scores of number of books read (for both sensitives and nonsensitives) is also skewed to the right. This shape for the population distribution also seems reasonable in light of what is being measured. A child cannot read fewer than zero books but once a child starts reading, it is easy to read a lot of books in a year.

Also note that the estimated population variances based on the two samples are dramatically different, 95.58 versus 584.25. This is another reason you would not want to go ahead with an ordinary *t* test.

However, suppose you do a square-root transformation on the scores (Table 11). Now both samples are much more like a normal curve; they have their middle scores bunched up in the middle (for example, for the Yes group, 6.00 and 6.71) and the more extreme (high and low) scores spread out a little from the mean (4.12 and 8.66). Also, the transformation seems reasonable in terms of the meaning of the numbers. The number of books read is meant as a measure of interest in things literary. Thus, the difference between 0 and 1 book is a much greater difference than the difference between 20 and 21 books.

Table 11 Square-Root Transformation of the Scores in Table 10

Highly Sensitive			
No		**Yes**	
X	$\sqrt{X}$	X	$\sqrt{X}$
0	0.00	17	4.12
3	1.73	36	6.00
10	3.16	45	6.71
22	4.69	75	8.66

Table 12 shows the *t* test analysis using the transformed scores.

Table 12 Figuring for a *t* Test for Independent Means Using Square-Root Transformed Scores for the Study of Books Read by Highly Sensitive versus Not Highly Sensitive Children (Fictional Data)

t needed for .05 significance level, $df = (4 - 1) + (4 - 1) = 6$, one-tailed $= -1.943$.

	Highly Sensitive	
	No	Yes
	0.00	4.12
	1.73	6.00
	3.16	6.71
	4.69	8.66
Σ:	9.58	25.49
$M =$	$9.58/4 = 2.40$	$25.49/4 = 6.37$
$S^2 =$	$12.03/3 = 4.01$	$10.56/3 = 3.52$
	$S^2_{\text{Pooled}} = 3.77$	
$S^2_M =$	$3.77/4 = .94$	$3.77/4 = .94$

$S^2_{\text{Difference}} = .94 + .94 = 1.88$

$S_{\text{Difference}} = \sqrt{1.88} = 1.37$

$t = (2.40 - 6.37)/1.37 = -2.90$

Decision: Reject the null hypothesis.

How are you doing?

1. What is a data transformation?
2. Why is it done?
3. When is this legitimate?
4. Consider the following distribution of scores: 4, 16, 25, 25, 25, 36, 64. (a) Are these scores roughly normally distributed? (b) Why? (c) Carry out a square-root transformation for these scores (that is, list the square-root transformed scores). (d) Are the square-root transformed scores roughly normally distributed? (e) Why?

Answers

1. A data transformation is when each score is changed following some rule (such as take the square root or log).
2. It is done to make the distribution more like a normal curve (or to make variances closer to equal across groups).
3. It is legitimate when it is done to all the scores, it is not done to make the results come out to fit the researcher's predictions, and the underlying meaning of the distance between scores is arbitrary.
4. (a) The scores are not roughly normally distributed. (b) They are skewed to the right. (c) Square root transformation: 2, 4, 5, 5, 5, 6, 8. (d) The square-root transformed scores are roughly normally distributed. (e) The middle scores are bunched in the middle and the extremes spread out evenly on both sides.

Rank-Order Tests

Another way of coping with non-normal distributions is to transform the scores to ranks. Suppose you have a sample with scores 4, 8, 12, and 64. This would be a rather surprising sample if the population was really normal. A **rank-order transformation** would change the scores to 1, 2, 3, and 4, the 1 referring to the lowest number in the group, the 2 to the second lowest, and so forth. A complication with a rank-order transformation occurs when you have two or more scores that are tied. The usual solution to ties is to give them each the average rank. For example, the scores 12, 81, 81, 107, and 154 would be ranked 1, 2.5, 2.5, 4, and 5.

Changing ordinary scores to ranks is a kind of data transformation. But unlike square-root transformations, with a rank-order transformation you aren't trying to get a normal distribution. The distribution you get from a rank-order transformation is rectangular, with equal numbers of scores (one) at each value. Ranks have the effect of spreading the scores out evenly.

There are special hypothesis-testing procedures that make use of rank-ordered data, called **rank-order tests.** They also have two other common names. You can transform scores from a population with any shaped distribution into ranks. Thus, these tests are sometimes called **distribution-free tests.** Also, the distribution of rank-order scores is known exactly rather than estimated. Thus, rank-order tests do not require estimating any parameters (population values). For example, there is no need to estimate a population variance because you can determine exactly what it will be if you know that ranks are involved. Hence, hypothesis-testing procedures based on ranks are also called **nonparametric tests.**

The ordinary hypothesis-testing procedures you have learned (*t* test and analysis of variance) are examples of **parametric tests.** Chi-square, like the rank-order tests, is considered a nonparametric test, but it is distribution-free only in the sense that no assumptions are made about the shape of the population distributions. However, the terms *distribution-free* and *nonparametric* are typically used interchangeably; the subtleties of differences between them are a matter of ongoing debate among statisticians.

Rank-order tests also have the advantage that they can be used where the actual scores in the study are themselves ranks—for example, a study comparing the class standings of two types of graduates.

Overview of Rank-Order Tests

Table 13 shows the name of the rank-order test that you would substitute for each of the parametric hypothesis-testing procedures you have learned. (Full procedures for using such tests are given in intermediate statistics texts.) For example, in a situation with three groups where you meet the assumptions, you would do a one-way analysis of variance. But if you don't meet the assumptions, you could use instead the rank-order version of the analysis of variance, the Kruskal-Wallis *H* test.

Next, we describe how such tests are done in a general way, including an example. However, we do not actually provide all the needed information for you to carry them out in practice. We introduce you to these techniques because you may see them used in articles you read and because their logic is the foundation of an alternative procedure that we do teach you to use (shortly). This alternative procedure does roughly the same thing as these rank-order tests and is closer to what you have already learned.

rank-order transformation Changing a set of scores to ranks (for example, so that the lowest score is rank 1, the next lowest rank 2, etc.).

rank-order test Hypothesis-testing procedure that uses rank-ordered scores.

distribution-free test Hypothesis-testing procedure making no assumptions about the shape of the populations; also called a *nonparametric test.*

nonparametric test Hypothesis-testing procedure making no assumptions about population parameters; also called a *distribution-free test.*

parametric test Ordinary hypothesis-testing procedure, such as a *t* test or an analysis of variance, that requires assumptions about the shape or other parameters (such as the variance) of the populations.

Table 13 Major Rank-Order Tests Corresponding to Major Parametric Tests

Ordinary Parametric Test	Corresponding Rank-Order Test
t test for dependent means	Wilcoxon signed-rank test
t test for independent means	Wilcoxon rank-sum test or Mann-Whitney *U* test
Analysis of variance	Kruskal-Wallis *H* test

Basic Logic of Rank-Order Tests

Consider a study involving an experimental group and a control group. (This is the kind of situation for which you would use a *t* test for independent means if all the assumptions were met.) If you wanted to use a rank-order test, you would first transform all the scores into ranks, ranking all the scores from lowest to highest, regardless of whether a score was in the experimental or the control group. If the two groups were scores randomly taken from a single population, there should be about equal amounts of high ranks and low ranks in each group. (That is, if the null hypothesis is true, the ranks in the two groups should not differ.) Because the distribution of ranks can be worked out exactly, statisticians can figure the exact probability of getting any particular division of ranks into two groups if in fact the two groups were randomly taken from identical distributions.

The way this actually works is that the researcher converts all the scores to ranks, adds up the total of the ranks in the group with the lower scores, and then compares this total to a cutoff from a special table of significance cutoffs for totals of ranks in this kind of situation. (Also, as with ordinary parametric tests, this can be done automatically with SPSS or other statistical software programs.)

An Example of a Rank-Order Test

Table 14 shows the transformation to ranks and the computation of the Wilcoxon rank-sum test for the kind of situation we have just described, using the books read by highly sensitive versus not highly sensitive children example. The logic is a little different from what you are used to, so be patient until we explain it.

Table 14 Figuring for a Wilcoxon Rank-Sum Test for the Study of Books Read by Highly Sensitive versus Not Highly Sensitive Children (Fictional Data)

Cutoff for significance: Maximum sum of ranks in the not highly sensitive group for significance at the .05 level, one-tailed (from a standard table) = 11.

Highly Sensitive			
No		Yes	
X	Rank	*X*	Rank
0	1	17	4
3	2	36	6
10	3	45	7
22	5	75	8
Σ:	11		

Comparison to cutoff: Sum of ranks of group predicted to have lower scores, 11, equals but does not exceed cutoff for significance.

Decision: Reject the null hypothesis.

Notice that we first set the significance cutoff, as you would in any hypothesis-testing procedure. (This cutoff is based on a table you don't have but is available in most intermediate statistics texts.)

The next step is to rank all the scores from lowest to highest, then add up the ranks in the group you expect to have the smaller total. You then compare the smaller total to the cutoff. If this smaller total is less than or equal to the cutoff, you reject the null hypothesis. In the example, the total of the ranks for the lower was actually equal to the cutoff, so the null hypothesis was rejected.

TIP FOR SUCCESS

Remember that when changing ordinary scores to ranks, all of the scores are ranked from lowest to highest, regardless of what group a score is in (for example, experimental vs. control).

Using Parametric Tests with Rank-Transformed Data

Two statisticians (Conover & Iman, 1981) have shown that instead of using the special procedures for rank-order tests, you get approximately the same results for the *t* test and one-way analysis of variance if you first transform the data into ranks and then just use all the usual *t* test or one-way analysis of variance procedures.

The result of this shortcut (using a parametric test with data transformed into ranks) will not be quite as accurate as either the ordinary parametric test or the rank-order test. It will not be as accurate because you are violating the assumption of normal population distributions. As we noted earlier, when you are using ranks, the population distribution is in fact rectangular (there are equal numbers—one—of each rank). Using this shortcut will also not be quite as accurate as the rank-order test. This is because the parametric test uses the *t* or *F* distribution instead of the special tables that rank-order tests use, which are based on exact probabilities of getting certain divisions of ranks. However, it turns out that, in practice, using an ordinary parametric test with ranks gives a result that is quite close to the true, accurate result you would get using the technically proper procedure.[5] Table 15 shows the figuring for an ordinary *t* test for independent means for the fictional sensitive children data, using each child's rank instead of the child's actual number of books read. Again we get a significant result. (In practice, carrying out an ordinary procedure like a *t* test with scores that have been transformed to ranks is least accurate with a very small sample like this. However, we used the small sample to keep the example simple.)

Comparison of Methods

We have considered two methods of carrying out hypothesis tests when samples appear to come from non-normal populations: data transformation and rank-order tests. How do you decide which to use?

Advantages and Disadvantages

Data transformations have the advantage of allowing you to use the familiar parametric techniques on the transformed scores. Transformations may also come closer to the true meaning of the underlying measurement. But transformations will not always work. For example, in an analysis of variance, there may not be any reasonable transformation

[5] If you want to be very accurate, for a *t* test or one-way analysis of variance, you can convert your result to what is called an *L* statistic and look it up on a chi-square table (Puri & Sen, 1985). The *L* statistic for a *t* test is $[(N-1)t^2]/[t^2 + (N-2)]$ and you use a chi-square distribution with $df = 1$. The *L* statistic for a one-way analysis of variance is $[(N-1)(df_{\text{Between}})F]/[(df_{\text{Between}})F + df_{\text{Within}}]$, and you use a chi-square distribution with $df = df_{\text{Between}}$. The *L* for the significance of a correlation is just $(N-1)r^2$ and you use the chi-square table for $df = 1$. It is especially important to use the *L* statistic when using rank-transformed scores for more advanced parametric procedures, such as factorial analysis of variance and multiple regression. Thomas, Nelson, and Thomas (1999) give fully worked-out examples.

Table 15 Figuring for a *t* Test for Independent Means Using Ranks Instead of Raw Scores for the Study of Books Read by Highly Sensitive versus Not Highly Sensitive Children (Fictional Data)

t needed for .05 significance level, $df = (4 - 1) + (4 - 1) = 6$, one-tailed $= -1.943$

	Highly Sensitive	
	No	Yes
	1	4
	2	6
	3	7
	5	8
Σ	11	25
$M =$	$11/4 = 2.75$	$25/4 = 6.25$
$S^2 =$	$8.75/3 = 2.92$	$8.75/3 = 2.92$
	$S^2_{\text{Pooled}} = 2.92$	
$S^2_M =$	$2.92/4 = .73$	$2.92/4 = .73$

$S^2_{\text{Difference}} = .73 + .73 = 1.46$

$S_{\text{Difference}} = \sqrt{1.46} = 1.21$

$t = (2.75 - 6.25)/1.21 = -2.89$

Decision: Reject the null hypothesis.

that makes the scores normal or have equal variances in all groups. Also, transformations may distort the scores in ways that lose the original meaning.

You can use rank-order methods regardless of the shape of the distributions of the original scores. Rank-order tests are, of course, especially appropriate when the original scores are ranks. They are especially useful when the scores do not clearly follow a simple numeric pattern (such as equal-interval), which some behavioral and social scientists think is a common situation. Furthermore, the logic of rank-order methods is simple and direct, requiring no elaborate construction of hypothetical distributions or estimated parameters.

However, rank-order methods are not as familiar to readers of research, and rank-order methods have not been developed for many complex situations. Another problem is that the simple logic of rank-order tests breaks down if there are many ties in ranks. Finally, like data transformation methods, rank-order methods distort the original data, losing information. For example, in the same sample, a difference between 6.1 and 6.2 could be one rank, but the difference between 3.4 and 5.8 might also be one rank.

Relative Risk of Type I and Type II Errors

How accurate are the various methods in terms of the 5% significance level really meaning that there is a 5% chance of incorrectly rejecting the null hypothesis (a Type I error)? And how do the different methods affect power?

When the assumptions for parametric tests are met, the parametric tests are as good as or better than any of the alternatives. This is true for protection against both Type I and Type II errors. This would be expected, because these are the conditions for which the parametric tests were designed.

However, when the assumptions for a parametric test are not met, the relative advantages of the possible alternative procedures we have considered (data transformation and rank-order tests) are not at all clear. In fact, the relative merits of the various procedures are topics of lively controversy, with many articles appearing in statistics-oriented journals every year.

How are you doing?

1. (a) What is a rank-order transformation? (b) Why is it done? (c) What is a rank-order test?
2. Transform the following scores to ranks: 5, 18, 3, 9, 2.
3. (a) If you wanted to use a standard rank-order test instead of a *t* test for independent means, what procedure would you use? (b) What are the steps of doing such a test?
4. (a) What happens if you change your scores to ranks and then figure an ordinary parametric test using the ranks? (b) Why will this not be quite as accurate, even assuming that the transformation to ranks is appropriate? (c) Why will this result not be quite as accurate using the standard rank-order test?
5. If conditions are not met for a parametric test, (a) what are the advantages and (b) disadvantages of data transformation over rank-order tests, and what are the (c) advantages and (d) disadvantages of rank-order tests over data transformation?

Answers

1. (a) A rank-order transformation is changing each score to its rank order (from lowest to highest) among all the scores.
(b) It is done to make the distribution a standard shape.
(c) A rank-order test is a special type of significance testing procedure designed for use with rank-ordered scores.

2. 5 = 3, 18 = 5, 3 = 2, 9 = 4, and 2 = 1. (That is, the ranks are 3, 5, 2, 4, and 1.)

3. (a) Wilcoxon rank-sum test or Mann-Whitney *U* test.
(b) Set the significance cutoff (based on a table) for the maximum sum of ranks for the group predicted to have the lower scores, change all scores to ranks (ignoring what group they are in), add up the ranks in the group predicted to have the lower scores, and then compare that total to the cutoff.

4. (a) If you change your scores to ranks and then figure an ordinary parametric test using the ranks, you get fairly similar results to doing the standard parametric test.
(b) The population distribution will be rectangular and not normal (an assumption for the *t* test).
(c) The rank-order test is based on knowing for sure the shape of the population distribution and using exact probabilities on that basis.

5. (a) You can use the familiar parametric methods, and the transformation may come closer to the true meaning of the underlying measurement.
(b) They will not always work and may distort the underlying meaning of the measurement.
(c) They can be used regardless of the distribution, and rank order may better reflect the true meaning of the measurement.
(d) They are often unfamiliar and have not been developed for many complex methods; also, ties in ranks (which are common) distort the accuracy of these tests.

Chi-Square Tests, Data Transformations, and Rank-Order Tests in Research Articles

Chi-Square Tests

In research articles, chi-square tests for goodness of fit usually include the frequencies in each category or cell, as well as the degrees of freedom, number of participants, the sample's chi-square, and significance level. For example, Black et al. (2009) reported their finding for the gender of characters on cereal boxes as follows: "[A] chi-square goodness-of-fit test was conducted on the total number of characters whose gender could be determined ($n = 1{,}386$). Seventy-two percent ($n = 996$) were male characters and 28% ($n = 390$) were female characters, which represents a significant disparity, $\chi^2\ (1) = 264.96$, $p < .001$" (p. 886).

Here is another example of a chi-square test for goodness of fit. Sandra Moriarty and Shu-Ling Everett (1994) did a study of television viewing in which graduate students actually went to 55 different homes and observed people watching television for 45-minute sessions. In one part of their results, they compared the number of people they observed who fell into one of four distinct categories:

> Flipping [very rapid channel changing], the category dominated by the most active type of behavior, occurred most frequently, in 33% of the sessions ($n = 18$). The grazing category [periods of browsing through channels] dominated 24% of the sessions ($n = 13$), and 22% were found to be in each of the continuous and stretch viewing categories ($n = 12$). These differences were not statistically significant ($\chi^2 = 1.79$, $df = 3$, $p > .05$). (p. 349)

Published reports of chi-square tests for independence provide the same basic chi-square information. For example, Durkin and Barber (2003) studied the relationship between playing computer games and positive development (such as being close to one's family, involved in activities, having positive mental health, and low disobedience to parents) among 16-year-old high school students in Michigan. As part of the study, the researchers tested whether male and female students differed in how often they played computer games. Students indicated how often they played computer games with a 7-point scale, from *never* (1) to *daily* (7). Here is how the researchers reported their results:

> The participants were categorized into three groups based on their frequency of play: "None" included participants who did not use computers at all, as well as those who used computers, but never for computer games; "Low" included participants who checked 2, 3, 4, or 5 for frequency of computer use to play computer games; and "High" included participants who checked 6 or 7 for frequency of computer game play. A chi-square test [for independence] indicated that males and females were not evenly distributed across these three categories [$\chi^2(2, N = 1043) = 62.39$, $p < .001$]. Girls were overrepresented among the nonusers, with a majority never playing computer games (50.6%), compared to 29.4% of boys who never played. Boys were more than twice as likely (23.8%) as girls (9.9%) to be in the high use group. A substantial number of both girls (39.4%) and boys (46.8%) were in the low use group. (p. 381)

You may be interested to read the researchers' conclusions from the overall study: "No evidence was obtained of negative outcomes among game players. On several measures—including . . . [all of the positive development outcomes mentioned earlier]—game players scored more favorably than did peers who never played computer games. It is concluded that computer games can be a positive feature of a healthy adolescence" (Durkin & Barber, 2002, p. 373). Researchers continue to examine the potential effects of playing video games, and there is an increasing focus on the content of games. As noted by Greitemeyer and Osswald (2010), ". . . there has been accumulating evidence that exposure to violent video games leads to increased aggressive behavior while decreasing prosocial behavior" (p. 212). However, there is also evidence that playing prosocial video games (that is, games that primarily involve helping another game character) can have positive effects. For example, based on the results of several studies in which university students played prosocial video games, Greitemeyer and Osswald concluded that ". . . participants who had played prosocial video games were more likely to help after a mishap, were more willing (and devoted more time) to assist in further experiments, and intervened more often in a harassment situation" (p. 211).

Black et al. (2009) reported the result of their chi-square test for independence of the association between the gender and age of characters on cereal boxes as follows: "The . . . hypothesis was that male characters would be more likely than female characters to be displayed as . . . adults. . . . A chi-square [analysis] tested the extent to which gender was associated with . . . age. . . . [F]emale characters were more likely than male characters to be depicted as children . . . , and male characters were more likely than female characters to be depicted as adults $[\chi^2(1) = 15.62, p < .001]$" (p. 886).

Data Transformations

Data transformations are usually mentioned in the Results section, just prior to the description of the analysis that uses the scores that were transformed. For example, Sugerman and Carey (2007) studied the relationship between students' alcohol intake and their use of strategies to control drinking. (Examples of strategies they studied were spacing drinks over time and alternating alcoholic and nonalcoholic drinks when drinking.) Prior to presenting the main results, the researchers noted:

> Summary statistics were generated to evaluate the distributions of variables and to identify problems with skew that might require transformations. To correct for nonnormality due to positive skew, we square-root transformed the following variables: average drinks per week, average BAC [blood alcohol content], and heaviest BAC. (p. 341)

Rank-Order Tests

Here is an example of rank-order tests reported by Schwitzgebel, Huang, and Zhou (2007). These researchers studied factors associated with how often people report dreaming in color. The participants in the study were 300 high school and university students in a central part of Eastern China. The students answered the question "Do you see colors in your dreams?" using response options of *never, rarely, occasionally, frequently,* and *very frequently.* The researchers noted at the start of the Results section of the article that "The data were treated as ranked and nonparametric" (p. 38). They went on to state that: "Respondents with a principally urban childhood . . . reported significantly more colored dreaming (median *occasionally*) than respondents raised in rural areas (median *rarely*) (Mann-Whitney, one-tailed, $p < .0001$)" (p. 40).

How often do *you* see colors in your dreams? Here is how a sample of 124 college students in Southern California answered that question: never, 4.7%; rarely, 14.3%; occasionally, 24.4%; frequently, 27.8%; very frequently, 28.7% (Schwitzgebel, 2003).

Learning Aids

Summary

1. Chi-square tests are used for hypothesis tests with *nominal variables.* A sample's chi-square statistic (χ^2) shows the amount of mismatch between expected and observed frequencies over several categories. It is figured by finding, for each category or combination of categories, the difference between observed frequency and expected frequency, squaring this difference (eliminating positive and negative signs), and dividing by the expected frequency (making the squared differences more proportionate to the numbers involved). The results are then added up for all the categories or combinations of categories. The distribution of the chi-square statistic is known and the cutoffs can be looked up in standard chi-square tables.
2. The chi-square test for goodness of fit is used to test hypotheses about whether a distribution of frequencies over two or more categories of a *single nominal variable* matches an expected distribution. (These expected frequencies are based, for example, on theory or on a distribution in another study or circumstance.) In this test, the expected frequencies are given in advance or are based on some expected percentages (such as equal percentages in all groups). The degrees of freedom are the number of categories minus 1.
3. The chi-square test for independence is used to test hypotheses about the relation between *two nominal variables*—that is, about whether the breakdown over the categories of one variable has the same proportional pattern in each of the categories of the other variable. The frequencies are set up in a contingency table, in which the two variables are crossed and the numbers in each combination are placed in each of the resulting cells. The frequency expected for a cell if the two variables are independent is the percentage of all the people in that cell's row multiplied by the total number of people in that cell's column. The degrees of freedom for the chi-square test for independence are the number of columns minus 1 multiplied by the number of rows minus 1.
4. Chi-square tests make no assumptions about normal distributions of their variables, but they do require that no individual be counted in more than one category or cell.
5. The estimated effect size for a chi-square test for independence (that is, the degree of association) for a 2 × 2 contingency table is the phi coefficient; for larger tables, Cramer's phi. Phi is the square root of the result of dividing your sample's chi-square by the number of persons. Cramer's phi is the square root of the result of dividing your sample's chi-square by the product of the number of persons multiplied by the degrees of freedom in the smaller side of the contingency table. These coefficients range from 0 to 1.
6. The *t* test, the analysis of variance, and other standard parametric tests all assume that populations follow a normal curve and have equal variances. When samples suggest that the populations are very far from normal (as when they are highly skewed or have outliers), using the ordinary procedures gives incorrect results.
7. One approach when the populations appear to be non-normal is to transform the scores, such as taking the square root or log of each score so that the distribution of the transformed scores appears to represent a normally distributed population. The ordinary hypothesis-testing procedures can then be used.
8. Another approach is to rank all of the scores in a study. Special rank-order tests (sometimes called nonparametric or distribution-free tests) use basic principles

of probability to determine the chance of the ranks being unevenly distributed across groups. However, in many situations, using the rank-transformed scores in an ordinary parametric test gives a good approximation.

9. Data transformations allow you to use the familiar parametric techniques but cannot always be used and may distort the meaning of the scores. You can use rank-order methods in almost any situation; they are especially appropriate with rank or similar data, and they have a straightforward conceptual foundation. But rank-order methods are not widely familiar and they have not been developed for many complex data analysis situations. As with data transformations, information may be lost or meaning distorted with rank-order methods.
10. Chi-square tests are reported in research articles using a standard format. For example, $\chi^2(3, N = 196) = 9.22, p < .05$. Research articles usually describe data transformations just prior to analyses using them. Rank-order methods are described much like any other kind of hypothesis test.

Key Terms

chi-square tests
chi-square test for goodness of fit
chi-square test for independence
observed frequency
expected frequency
chi-square statistic (χ^2)
chi-square distribution
chi-square table
contingency table
independence
cell
phi coefficient (ϕ)
Cramer's phi
data transformation
square-root transformation
rank-order transformation
rank-order tests
distribution-free tests
nonparametric tests
parametric tests

Chi-Square Test for Goodness of Fit

The expected distribution (from previous years) on an exam roughly follows a normal curve in which the highest scoring 2.5% of the students get As; the next highest scoring 14%, Bs; the next 67%, Cs; the next 14%, Ds; and the lowest 2.5%, Fs. A class takes a test using a new grading system and 10 get As, 34 get Bs, 140 get Cs, 10 get Ds, and 6 get Fs. Use the steps of hypothesis testing to decide whether the new system produces a different distribution of grades (using the .01 level).

Answer

Table 16 shows the observed and expected frequencies and the figuring for the chi-square test.

❶ **Restate the question as a research hypothesis and a null hypothesis about the populations.** There are two populations:

Population 1: Students like those graded with the new system.
Population 2: Students like those graded with the old system.

The research hypothesis is that the populations are different; the null hypothesis is that the populations are the same.

Table 16 Figuring for Chi-Square Test for Goodness of Fit Example Worked-Out Problem

Grade	Observed Ⓐ	Expected Ⓑ
A	10	5 (2.5% × 200)
B	34	28 (14.0% × 200)
C	140	134 (67.0% × 200)
D	10	28 (14.0% × 200)
F	6	5 (2.5% × 200)

Degrees of freedom $= N_{\text{Categories}} - 1 = 5 - 1 = 4$ ❷

Chi-square needed, $df = 4$, .01 level: 13.277 ❸

$$\chi^2 = \Sigma\frac{(O-E)^2}{E} = \frac{(10-5)^2}{5} + \frac{(34-28)^2}{28} + \frac{(140-134)^2}{134} + \frac{(10-28)^2}{28} + \frac{(6-5)^2}{5}$$

$$= \frac{5^2}{5} + \frac{6^2}{28} + \frac{6^2}{134} + \frac{\overset{\text{Ⓒ}}{-18^2}}{28} + \frac{1^2}{5}$$

$$= \frac{25}{5} + \frac{36}{28} + \frac{\overset{\text{Ⓓ}}{36}}{134} + \frac{324}{28} + \frac{1}{5}$$

$$= 5 + 1.29 + .27 + \underset{\text{Ⓔ}}{11.57} + .20 = \overset{\text{Ⓕ}}{18.33}.$$

Decision: Reject the null hypothesis. ❺

❷ **Determine the characteristics of the comparison distribution.** The comparison distribution is a chi-square distribution with 4 degrees of freedom ($df = N_{\text{Categories}} - 1 = 5 - 1 = 4$).

❸ **Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected.** Using the .01 level and $df = 4$, Table 4 of the appendix "Tables" shows a needed chi-square of 13.277.

❹ **Determine your sample's score on the comparison distribution.** As shown in Table 16, this comes out to 18.33.

❺ **Decide whether to reject the null hypothesis.** The sample's chi square of 18.33 is more extreme than the needed chi-square of 13.277. Thus, you can reject the null hypothesis and conclude that the populations are different; the new grading system produces a different distribution of grades than the previous one.

Chi-Square Test for Independence

Steil and Hay (1997) conducted a survey of professionals (lawyers, doctors, bankers, and the like) regarding the people they compare themselves to when they think about their job situation (salary, benefits, responsibility, status, etc.). One question of special interest was how much professionals compare themselves to people of their own sex, the opposite sex, or both. Here are the results:

	Participant Gender	
	Men	Women
Comparison		
Same sex	29	17
Opposite sex	4	14
Both sexes	26	28

Table 17 Figuring for Chi-Square Test for Independence Example Worked-Out Problem

		Participant Gender		
		Men Ⓐ	**Women** Ⓑ	**Total**
Comparison	Same sex	29 (23)	17 (23)	46 (39.0%)
	Opposite sex	4 (9)	14 (9)	18 (15.3%)
	Both sexes	26 (27)	28 (27)	54 (45.8%)
	Total	59	59	118

$df = (N_{\text{Columns}} - 1)(N_{\text{Rows}} - 1) = (2 - 1)(3 - 1) = (1)(2) = 2.$ ❷

Chi-square needed, $df = 2$, .05 level: 5.992. ❸

$$\chi^2 = \Sigma\frac{(O - E)^2}{E} = \frac{(29 - 23)^2}{23} + \frac{(17 - 23)^2}{23} + \frac{(4 - 9)^2}{9} + \frac{(14 - 9)^2}{9} + \frac{(26 - 27)^2}{27} + \frac{(28 - 27)^2}{27}$$

$$= \frac{6^2}{23} + \frac{-6^2}{23} + \frac{-5^2}{9} + \frac{5^2}{9}\textcircled{C} + \frac{-1^2}{27} + \frac{1^2}{27}$$

$$= \frac{36}{23} + \frac{36}{23} + \frac{25}{9} + \frac{25}{9}\textcircled{D}\textcircled{E} + \frac{1}{27} + \frac{1}{27}$$

$$= 1.57 + 1.57 + 2.78 + 2.78 + .04 + .04 = 8.78. \textcircled{F}$$

Decision: Reject the null hypothesis. ❺

Note: Data from Stell and Hay (1997). The chi-square computed here (8.78) is slightly different from that reported in their article (8.76) due to rounding error.

Use the steps of hypothesis testing to decide if the researchers can conclude that the gender of who people compare themselves to is different depending on their own gender (use the .05 level).

Answer

Table 17 shows the figuring for the chi-square test.

❶ **Restate the question as a null hypothesis and a research hypothesis about the populations.** There are two populations:

Population 1: Professionals like those surveyed.
Population 2: Professionals for whom own sex is independent of the sex of those to whom they compare their job situations.

The null hypothesis is that the two populations are the same, that, in general, professional men and women do not differ in the sex of those to whom they compare their job situations. The research hypothesis is that the populations are not the same, that professionals like those surveyed are unlike the hypothetical population in which men and women do not differ in the sex of those to whom they compare their job situations.

❷ **Determine the characteristics of the comparison distribution.** This is a chi-square distribution with 2 degrees of freedom.

❸ **Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected.** Using the .05 level and 2 degrees of freedom, the needed chi-square for significance is 5.992.

❹ **Determine your sample's score on the comparison distribution.** As shown in Table 17, this comes out to 8.78.

❺ **Decide whether to reject the null hypothesis.** The chi square of 8.78 is larger than the cutoff of 5.992. Thus you can reject the null hypothesis: The gender of the people with whom professionals compare their job situations is likely to be different for men and women.

Effect Size for a 2 × 2 Chi-Square Test for Independence

Figure the effect size for a study with 85 participants and a chi-square of 14.41.

Answer

$$\phi = \sqrt{\chi^2/N} = \sqrt{14.41/85} = \sqrt{.170} = .41.$$

Effect Size for a Chi-Square Test for Independence with a Contingency Table Greater Than 2 × 2

Figure the effect size for the Steil and Hay (1997) example.

Answer

$$\text{Cramer's } \phi = \sqrt{\frac{\chi^2}{(N)(df_{\text{Smaller}})}} = \sqrt{\frac{8.78}{(118)(1)}} = \sqrt{.074} = .27.$$

Outline for Writing Essays for a Chi-Square Test for Goodness of Fit

1. Explain that chi-square tests are used for hypothesis testing with nominal variables. The chi-square test for goodness of fit is used to test hypotheses about whether a distribution of frequencies over two or more categories of a single nominal variable matches an expected distribution. Be sure to explain the meaning of the research hypothesis and the null hypothesis in this situation.
2. Describe the core logic of hypothesis testing in this situation. Be sure to mention that the hypothesis testing involves comparing observed frequencies (that is, frequencies found in the actual study) with expected frequencies (that is, frequencies that you would expect based on a particular theory or the results of previous research studies). The size of the discrepancy between the observed and expected frequencies determines whether the null hypothesis can be rejected.
3. Explain that the comparison distribution in this situation is a chi-square distribution. Be sure to mention that the shape of the chi-square distribution depends on the number of degrees of freedom. Describe how to determine the degrees of freedom and the cutoff chi-square value.
4. Describe how to figure the chi-square value for the sample. The key idea is to get a single number that indicates the overall discrepancy between what was found in the study and what would be expected based on some null hypothesis idea (such as the groups all being equal). To get this number you figure, for each group, the difference between the observed frequency and the expected frequency, square it (because otherwise the sign of the differences would cancel each other out when you added them up), and divide the squared difference by the expected frequency (to adjust for the size of the numbers involved). You then add up all of the adjusted squared differences to get an overall number. (This should all be explained using the numbers in the study as an example.)
5. Explain how and why the scores from Steps ❸ and ❹ of the hypothesis-testing process are compared. Explain the meaning of the result of this comparison with regard to the specific research and null hypotheses being tested.

Outline for Writing Essays for a Chi-Square Test for Independence

Follow the preceding general outline for the chi-square test for goodness of fit, noting that the chi-square test for independence is used to test hypotheses about the relation between two nominal variables. Using the actual numbers in your study as examples, be sure also to explain the concept of independence and how and why you figure the expected frequency for each cell (in terms of the cells in each column having the same proportions of the column total as the cell's row total is a proportion of the overall total).

Square-Root Transformation

Carry out a square-root transformation on the following scores from a study with three groups:

Group A	Group B	Group C
15	21	18
4	16	19
12	49	11
14	17	22

Answer

Group A	Group B	Group C
3.88	4.58	4.24
2.00	4.00	4.36
3.46	7.00	3.32
3.74	4.12	4.69

Rank-Order Transformation

Carry out a rank-order transformation on the same scores you used for the square-root transformation above.

Answer

Group A	Group B	Group C
5	10	8
1	6	9
3	12	2
4	7	11

These problems involve figuring. Most real-life statistics problems are done on a computer with special statistical software. Even if you have such software, do these problems by hand to ingrain the method in your mind. To learn how to use a computer to solve statistics problems like those in this chapter, refer to the "Using SPSS" section at the end of this chapter.

All data are fictional unless an actual citation is given.

Set I (for answers, see the end of this chapter)

1. Carry out a chi-square test for goodness of fit for each of the following (use the .05 level for each):

(a)	Category	Expected	Observed
	A	20%	19
	B	20%	11
	C	40%	10
	D	10%	5
	E	10%	5

(b)	Category	Expected	Observed
	I	30%	100
	II	50%	100
	III	20%	100

(c)		Number in	
	Category	the Past	Observed
	1	100	38
	2	300	124
	3	50	22
	4	50	16

2. A director of a social service agency is planning to hire temporary staff to assist with intake. In making plans, the director needs to know whether there is any difference in the use of the agency at different seasons of the year. Last year there were 28 new clients in the winter, 33 in the spring, 16 in the summer, and 51 in the fall. On the basis of last year's data, should the director conclude that season makes a difference? (Use the .05 level.) (a) Carry out the five steps of hypothesis testing for a chi-square test for goodness of fit. (b) Explain your answer to a person who has never taken a course in statistics.
3. Carry out a chi-square test for independence for each of the following contingency tables (use the .01 level). Also, figure the effect size for each.

(a)

10	16
16	10

(b)

100	106
106	100

(c)

100	160
160	100

(d)

10	16	10
16	10	10

(e)

10	16	16
16	10	16

(f)

10	16	10
16	10	16

4. A political analyst is interested in whether the community in which a person lives is related to that person's opinion on an upcoming water conservation ballot initiative. The analyst surveys 90 people by phone. The results are shown in the following table. Is opinion related to community at the .05 level? (a) Carry out the five steps of hypothesis testing. (b) Compute Cramer's phi and power. (c) Explain your answers to (a) and (b) to a person who has never taken a course in statistics.

	Community A	**Community B**	**Community C**
For	12	6	3
Against	18	3	15
No opinion	12	9	12

5. Figure the effect size for the following studies:

	N	**Chi-Square**	**Design**
(a)	100	16	2 × 2
(b)	100	16	2 × 5
(c)	100	16	3 × 3
(d)	100	8	2 × 2
(e)	200	16	2 × 2

6. What is the power of the following planned studies using a chi-square test for independence with $p < .05$?

	Predicted Effect Size	**Design**	***N***
(a)	small	2 × 2	25
(b)	medium	2 × 2	25
(c)	small	2 × 2	50
(d)	small	2 × 3	25
(e)	small	3 × 3	25
(f)	small	2 × 5	25

7. About how many participants do you need for 80% power in each of the following planned studies using a chi-square test for independence with $p < .05$?

	Predicted Effect Size	**Design**
(a)	medium	2 × 2
(b)	large	2 × 2
(c)	medium	2 × 5
(d)	medium	3 × 3
(e)	large	2 × 3

8. Lydon, Pierce, and O'Regan (1997) conducted a study that compared long-distance to local dating relationships. The researchers first administered questionnaires to a group of students one month prior to their leaving home to begin their first semester at McGill University (Time 1). Some of these students had dating partners who lived in the McGill area; others had dating partners who lived a long way from McGill. The researchers contacted the participants again late in the fall semester, asking them about the current status of their original dating relationships (Time 2). Here is how they reported their results:

   > Of the 69 participants . . . 55 were involved in long-distance relationships, and 14 were in local relationships (dating partner living within 200 km of them). Consistent with our predictions, 12 of the 14 local relationships were still intact at Time 2 (86%), whereas only 28 of the 55 long-distance relationships were still intact (51%), $\chi^2(1, N = 69) = 5.55, p < .02$. (p. 108)

   (a) Figure the chi-square yourself (your results should be the same, within rounding error).
   (b) Figure the effect size. (c) Explain the results to parts (a) and (b) to a person who has never had a course in statistics.
9. For each of the following distributions, make a square-root transformation:
   (a) 16, 4, 9, 25, 36
   (b) 35, 14.3, 13, 12.9, 18
10. A researcher compares the typical family size in 10 cultures, five from Language Group A and five from Language Group B. The figures for the Group A cultures are 1.2, 2.5, 4.3, 3.8, and 7.2. The figures for the Group B cultures are 2.1, 9.2, 5.7, 6.7, and 4.8. Based on these 10 cultures, does typical family size differ in cultures with different language groups? Use the .05 level. (a) Carry out a *t* test for independent means using the actual scores. (b) Carry out a square-root transformation (to keep things simple, round off the transformed scores to one decimal place). (c) Carry out a *t* test for independent means using the transformed scores. (d) Explain what you have done and why to someone who is familiar with the *t* test for independent means but not with data transformation.
11. A researcher randomly assigns participants to watch one of three kinds of films: one that tends to make people sad, one that tends to make people angry, and one that tends to make people exuberant. The participants are then asked to rate a series of photos of individuals on how honest they appear. The ratings for the sad-film group were 201, 523, and 614; the ratings for the angry-film group were 136, 340, and 301; and the ratings for the exuberant-film group were 838, 911, and 1,007. (a) Carry out an analysis of variance using the actual scores (use $p < .01$). (b) Carry out a square-root transformation of the scores (to keep things simple, round off the transformed scores to one decimal place). (c) Carry out an analysis of variance using the transformed scores. (d) Explain what you have done and why to someone who is familiar with analysis of variance but not with data transformation.
12. Miller (1997) conducted a study of commitment to a romantic relationship and how much attention a person pays to attractive alternatives. In this study, participants were shown a set of slides of attractive individuals. At the start of the Results section, Miller notes, "The self-reports on the Attentiveness to Alternative Index and the time spent actually inspecting the attractive opposite-sex slides . . . were positively skewed, so logarithmic transformations of the data were performed" (p. 760). Explain what is being described here (and why it is being done) to a person who understands ordinary parametric statistics but has never heard of data transformations.

Set II

13. Carry out a chi-square test for goodness of fit for each of the following (use the .01 level for each):

(a)	Category	Expected	Observed
	1	2%	5
	2	14%	15
	3	34%	90
	4	34%	120
	5	14%	50
	6	2%	20

(b)		Proportion	
	Category	Expected	Observed
	A	1/3	10
	B	1/6	10
	C	1/2	10

14. A researcher wants to be sure that the sample in her study is not unrepresentative of the distribution of ethnic groups in her community. Her sample includes 300 whites, 80 African Americans, 100 Latinos, 40 Asians, and 80 others. In her community, according to census records, there are 48% whites, 12% African Americans, 18% Latinos, 9% Asians, and 13% others. Is her sample unrepresentative of the population in her community? (Use the .05 level.) (a) Carry out the steps of hypothesis testing for a chi-square test for goodness of fit. (b) Explain your answer to a person who has never taken a course in statistics.

15. Carry out a chi-square test for independence for each of the following contingency tables (use the .05 level). Also, figure the effect size for each contingency table.

(a)

0	18
18	0

(b)

0	0	18
9	9	0

(c)

0	0	9	9
9	9	0	0

(d)

20	40
0	40

16. The following results are from a survey of a sample of people buying ballet tickets, laid out according to the type of seat they purchased and how regularly they attend. Is there a significant relation? (Use the .05 level.) (a) Carry out the steps of hypothesis testing. (b) Figure the effect size. (c) Explain your answer to parts (a) and (b) to a person who has never taken a course in statistics.

		Attendance	
		Regular	*Occasional*
Seating Category	*Orchestra*	20	80
	Dress circle	20	20
	Balcony	40	80

17. Figure the effect size for the following studies:

	N	Chi-Square	Design
(a)	40	10	2 × 2
(b)	400	10	2 × 2
(c)	40	10	4 × 4
(d)	400	10	4 × 4
(e)	40	20	2 × 2

18. What is the power of the following planned studies, using a chi-square test for independence with $p < .05$?

	Predicted Effect Size	Design	*N*
(a)	medium	2 × 2	100
(b)	medium	2 × 3	100
(c)	large	2 × 2	100
(d)	medium	2 × 2	200
(e)	medium	2 × 3	50
(f)	small	3 × 3	25

19. About how many participants do you need for 80% power in each of the following planned studies, using a chi-square test for independence with $p < .05$?

	Predicted Effect Size	Design
(a)	small	2 × 2
(b)	medium	2 × 2
(c)	large	2 × 2
(d)	small	3 × 3
(e)	medium	3 × 3
(f)	large	3 × 3

20. Everett, Price, Bedell, and Telljohann (1997) mailed a survey to a random sample of physicians. Half were offered $1 if they would return the questionnaire (this was the experimental group); the other half served as a control group. The point of the study was to see if even a small incentive would increase the return rate for physician surveys. Everett et al. report their results as follows:

> Of the 300 surveys mailed to the experimental group, 39 were undeliverable, 2 were returned uncompleted, and 164 were returned completed. Thus, the response rate for the experimental group was 63% $[164/(300 - 39) = .63]$. Of the 300 surveys mailed to the control group, 40 were undeliverable, and 118 were returned completed. Thus, the response rate for the control group was 45% $[118/(300 - 40) = .45]$. A chi-square test comparing the response rates for the experimental and control groups found the $1 incentive had a statistically significantly improved response rate over the control group $[\chi^2(1, N = 521) = 16.0, p < .001]$.

(a) Figure the chi-square yourself (your results should be the same, within rounding error). (b) Figure the effect size. (c) Explain the results to parts (a) and (b) to a person who has never had a course in statistics.

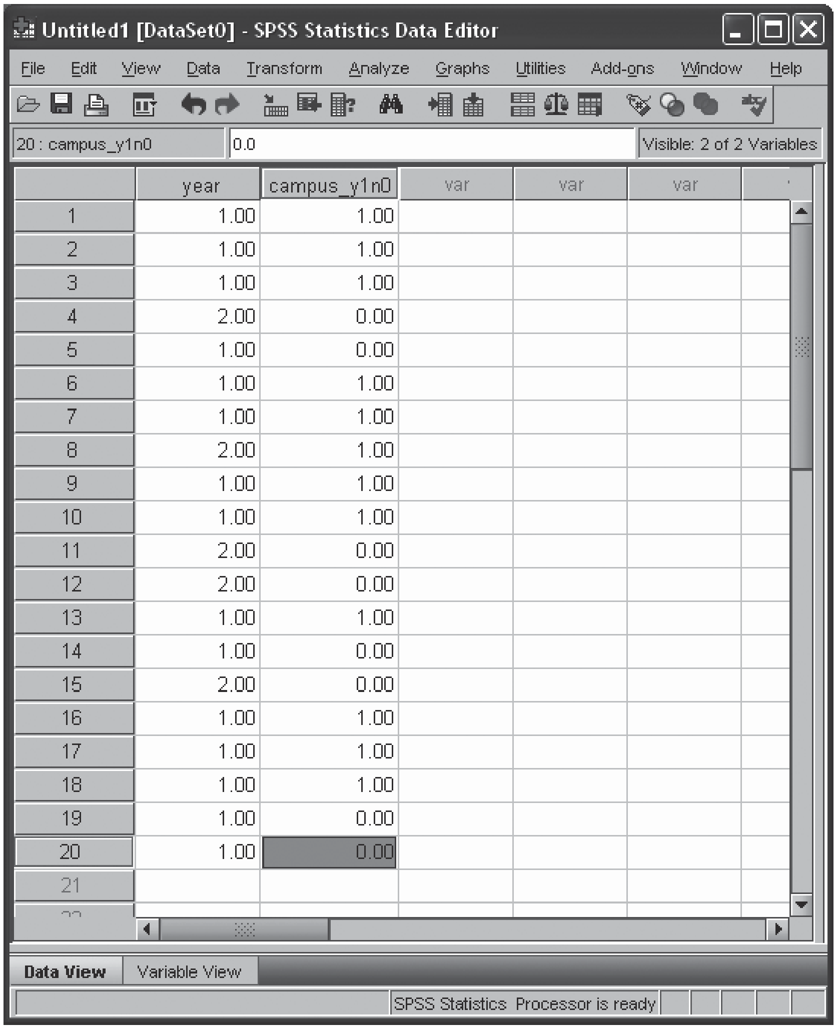
21. For each of the following distributions, make a square-root transformation:
    (a) 100, 1, 64, 81, 121
    (b) 45, 30, 17.4, 16.8, 47
22. A study compares performance on a novel task for fifth-grade students who do the task either alone, in the presence of a stranger, or in the presence of a friend. The scores for the students in the alone condition are 1, 1, and 0; the scores of the participants in the stranger condition are 2, 6, and 1; and the scores for those in the friend condition are 3, 9, and 10. (a) Carry out an analysis of variance using the actual scores ($p < .05$). (b) Carry out a square-root transformation of the scores (to keep things simple, round off the transformed scores to one decimal place). (c) Carry out an analysis of variance using the transformed difference scores. (d) Explain what you have done and why to someone who is familiar with analysis of variance but not with data transformation.
23. A researcher conducted an experiment organized around a major televised address by the U.S. president. Immediately after the address, three participants were randomly assigned to listen to the commentaries provided by the television networks' political commentators. The other three were assigned to spend the same time with the television off, reflecting quietly about the speech. Participants in both groups then completed a questionnaire that assessed how much of the content of the speech they remembered accurately. The group that heard the commentators had scores of 4, 0, and 1. The group that reflected quietly had scores of 9, 3, and 8. Did hearing the commentary affect memory? Use the .05 level, one-tailed, predicting higher scores for the reflected-quietly group. (a) Carry out a *t* test for independent means using the actual scores. (b) Carry out a square-root transformation (to keep things simple, round off the transformed scores to one decimal place). (c) Carry out a *t* test for independent means using the transformed scores. (d) Explain what you have done and why to someone who is familiar with the *t* test for independent means but not with data transformation.
24. As part of a larger study, Betsch, Plessner, Schwieren, and Gutig (2001) manipulated the attention to information presented in TV ads and then gave participants questions about the content of the ads as a check on the success of their manipulation. They reported:

    > Participants who were instructed to attend to the ads answered 51.5% . . . of the questions correctly. In the other condition, only 41.1% of questions were answered correctly. This difference is significant according to the Mann-Whitney U test, $U(84) = 2317.0$, $p < .01$. This shows that the attention manipulation was effective. (p. 248)

    Explain the general idea of what these researchers are doing (and why they didn't use an ordinary *t* test) to a person who is familiar with the *t* test but not with rank-order tests.

The ☞ in the steps below indicates a mouse click. (We used SPSS version 17.0 for Windows to carry out these analyses. The steps and output may be slightly different for other versions of SPSS.)

It is easier to learn the SPSS steps for chi-square tests using actual numbers. As an example, imagine you are a student in a class of 20 students and want to determine whether the students in the class are split equally among first- and second-year students


Untitled1 [DataSet0] - SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
20 : campus_y1n0
0.0
Visible: 2 of 2 Variables
var
var
var
Data View
Variable View
SPSS Statistics Processor is ready

	year	campus_y1n0
1	1.00	1.00
2	1.00	1.00
3	1.00	1.00
4	2.00	0.00
5	1.00	0.00
6	1.00	1.00
7	1.00	1.00
8	2.00	1.00
9	1.00	1.00
10	1.00	1.00
11	2.00	0.00
12	2.00	0.00
13	1.00	1.00
14	1.00	0.00
15	2.00	0.00
16	1.00	1.00
17	1.00	1.00
18	1.00	1.00
19	1.00	0.00
20	1.00	0.00
21		

Figure 10 SPSS data editor window for a fictional study examining the distribution of first- and second-year college students for a particular class and whether the students live on campus.

(the class is not open to students in other year groups). We will use a chi-square test for goodness of fit to answer this question. Each student's year in college is shown in the first column of Figure 10.

Chi-Square Test for Goodness of Fit

❶ Enter the scores into SPSS. As shown in Figure 10, the score for each person is listed in a separate row. We labeled the variable "year." (For now, you can ignore the "campus_y1n0" variable.)

❷ ☞ *Analyze.*

❸ ☞ *Nonparametric Tests,* ☞ *Chi-Square.*

❹ ☞ on the variable called "year" and then ☞ the arrow next to the box labeled "Test Variable List." This tells SPSS that the chi-square test for goodness of fit should be carried out on the scores for the nominal variable called "year." Notice in the

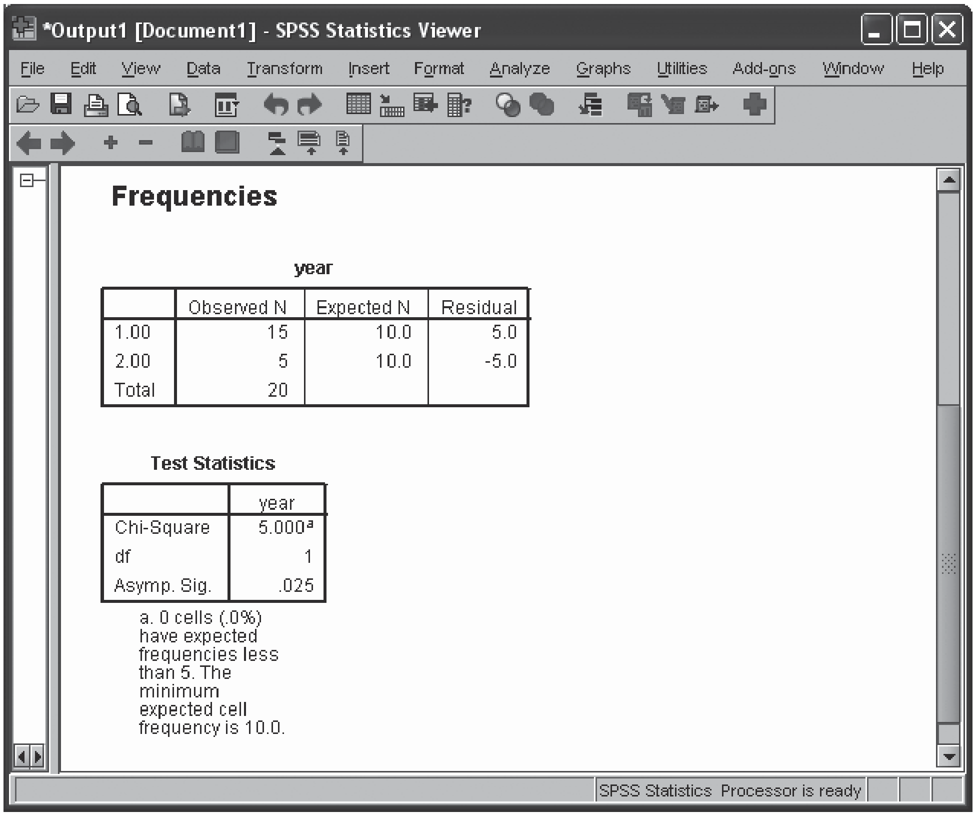
"Expected Values" box that the option "All categories equal" is selected by default. This means that SPSS will carry out the chi-square to compare the observed frequency distribution in your sample with an expected frequency distribution based on an equal spread of scores across the categories. If you wish to use a different expected frequency distribution, select the "Values" option and enter the appropriate expected values.

⑤ *OK.* Your SPSS output window should look like Figure 11.

The first table in the SPSS output gives the observed frequencies, the expected frequencies, and the difference between the observed and expected frequencies (in the "Residual" column). The second table gives the value of chi-square, the degrees of freedom, and the exact significance level. The significance level of .025 (for the chi-square value of 5.00) is less than our .05 cutoff, which means that you can reject the null hypothesis. Thus, you can conclude that the students in the class are not split equally among first- and second-year students.

Chi-Square Test for Independence

Using the same example as for the chi-square test for goodness of fit, let's suppose you are interested in whether the distribution of first- and second-year students in the class is different for students who live on campus versus those who do not live on campus.



Frequencies

year

	Observed N	Expected N	Residual
1.00	15	10.0	5.0
2.00	5	10.0	-5.0
Total	20		

Test Statistics

	year
Chi-Square	5.000[a]
df	1
Asymp. Sig.	.025

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 10.0.

SPSS Statistics Processor is ready

Figure 11 SPSS output window for a chi-square test for goodness of fit for a fictional study examining the distribution of first- and second-year college students for a particular class.

To answer this question (which involves *two nominal variables*), we will use a chi-square test for independence.

❶ Enter the scores into SPSS. As shown in Figure 10, the score for each person is listed in a separate row. We labeled the variables "year" and "campus_y1n0." (We assigned students who live on campus a value of "1" and students who don't live on campus a value of "0.")

❷ 🖱 *Analyze.*

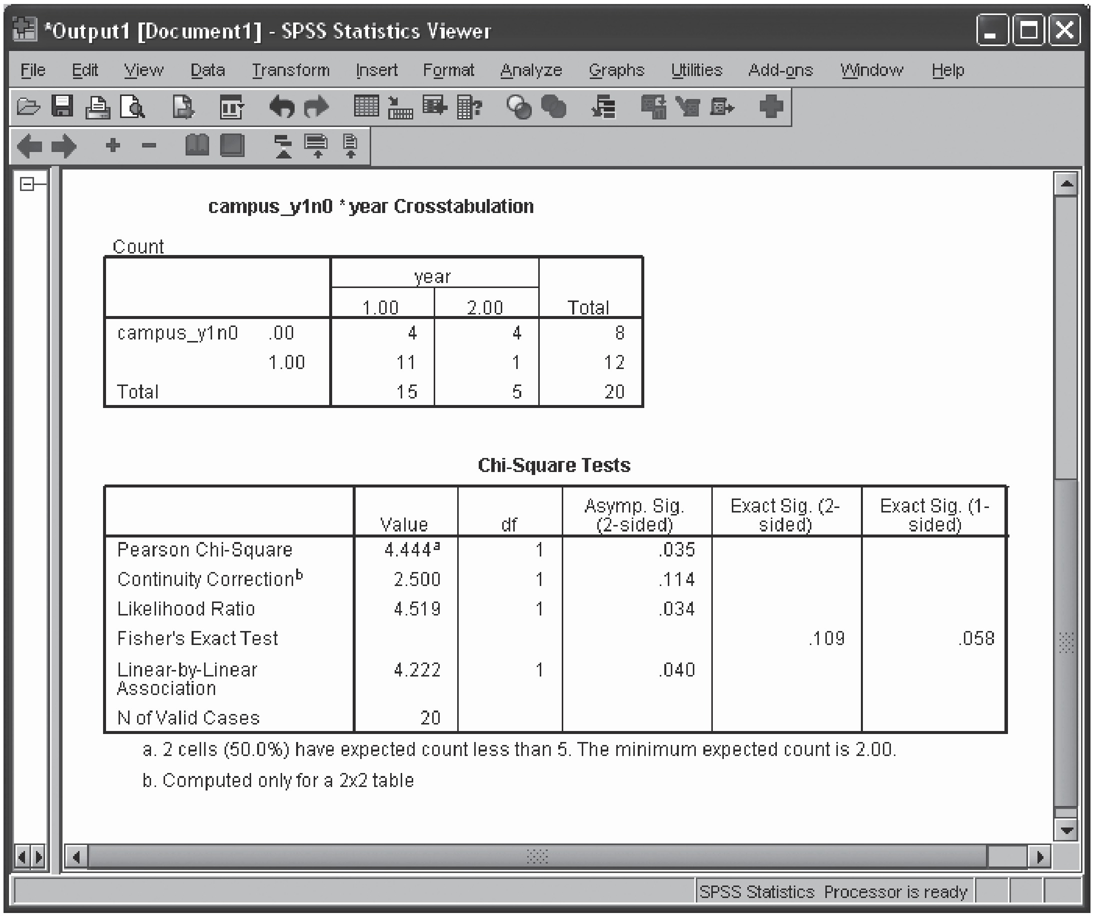
❸ 🖱 *Descriptive Statistics,* 🖱 *Crosstabs.*

❹ 🖱 on the variable called "campus_y1n0" and then 🖱 the arrow next to the box labeled "Row(s)." 🖱 on the variable called "year" and then 🖱 the arrow next to the box labeled "Column(s)." (It doesn't matter which variable is assigned to rows and which is assigned to columns; the result will be the same.)

❺ 🖱 *Statistics.* 🖱 the box labeled *Chi-square* (this checks the box). 🖱 *Continue.*

❻ 🖱 *OK.* Your SPSS output window should look like Figure 12.

The first table in the SPSS output (which is not shown in Figure 12) gives the number of individuals for each variable and whether there are any missing scores.


*Output1 [Document1] - SPSS Statistics Viewer
File
Edit
View
Data
Transform
Insert
Format
Analyze
Graphs
Utilities
Add-ons
Window
Help
SPSS Statistics Processor is ready

campus_y1n0 * year Crosstabulation

Count

		year		Total
		1.00	2.00	
campus_y1n0	.00	4	4	8
	1.00	11	1	12
Total		15	5	20

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	4.444[a]	1	.035		
Continuity Correction[b]	2.500	1	.114		
Likelihood Ratio	4.519	1	.034		
Fisher's Exact Test				.109	.058
Linear-by-Linear Association	4.222	1	.040		
N of Valid Cases	20				

a. 2 cells (50.0%) have expected count less than 5. The minimum expected count is 2.00.

b. Computed only for a 2x2 table

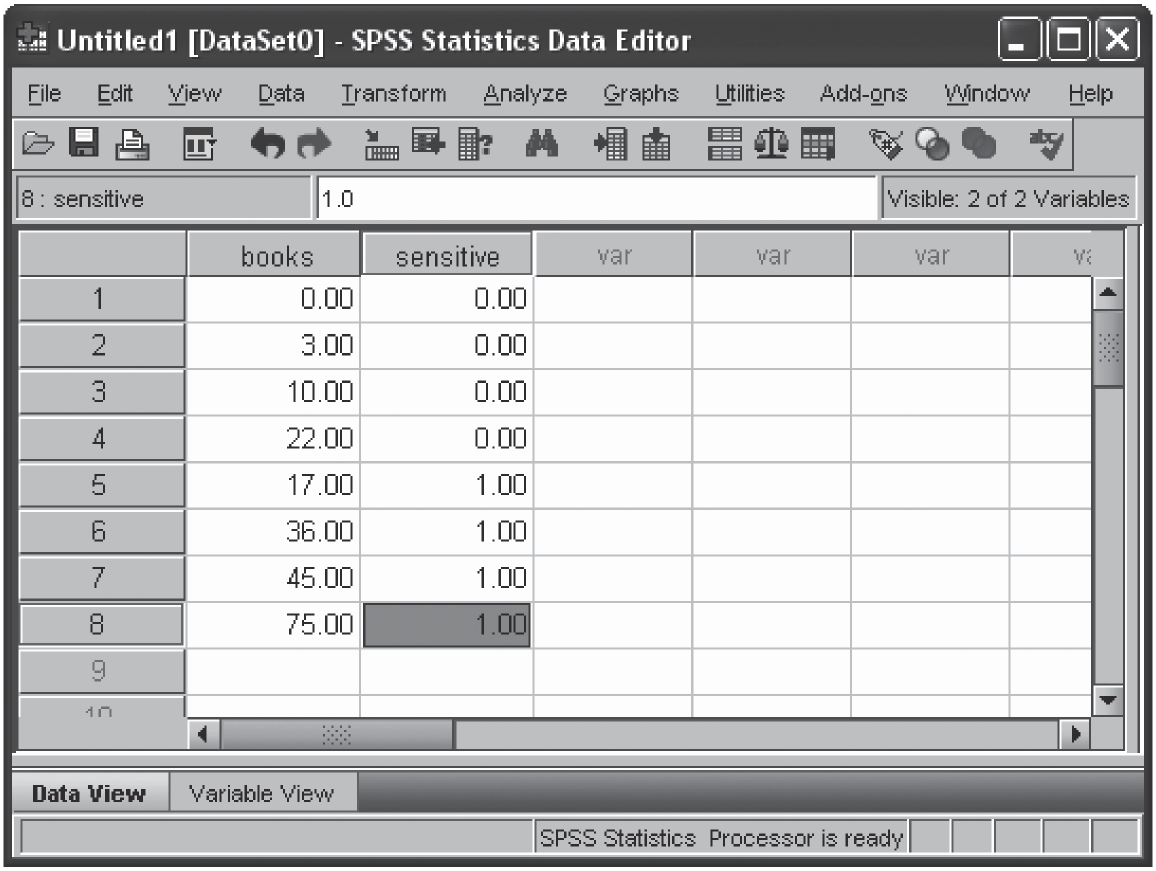
Figure 12 SPSS output window for a chi-square test for independence for a fictional study examining the distribution of first- and second-year college students for a particular class and whether the students live on campus.

The second table (labeled "campus_y1n0 * year Crosstabulation") gives the contingency table of observed values for the two nominal variables ("campus_y1n0" and "year"). The third table (labeled "Chi-Square Tests") shows the actual result of the chi-square test for independence, as well as the results of other tests. The results of the chi-square test for independence are provided in the first row (labeled "Pearson Chi-Square"), which shows the chi-square value, the degrees of freedom, and the exact significance level. The significance level of .035 (for the chi-square value of 4.444) is less than our .05 cutoff. Thus, you can reject the null hypothesis. This means you can conclude that the distribution of first- and second-year students is different according to whether students live on campus or off campus.

Data Transformations

We will use the example of the square-root transformation of the scores from the study comparing highly sensitive and not highly sensitive children on the number of books read in the past year (see Tables 10 and 11).

❶ Enter the scores into SPSS. As shown in Figure 13, the score for each child is shown in the "books" column. The scores in the "sensitive" column show whether the child was not highly sensitive (a score of "0") or highly sensitive



8 : sensitive | 1.0 | Visible: 2 of 2 Variables

	books	sensitive	var	var	var	va
1	0.00	0.00				
2	3.00	0.00				
3	10.00	0.00				
4	22.00	0.00				
5	17.00	1.00				
6	36.00	1.00				
7	45.00	1.00				
8	75.00	1.00				
9						
10						

SPSS Statistics Processor is ready

Figure 13 SPSS data editor window for the fictional study comparing highly sensitive and not highly sensitive children on the number of books read in the past year.

(a score of "1"). Although the "sensitive" scores aren't needed for the data transformation, they are important for figuring a *t* test for independent means on the transformed scores (see Table 12).

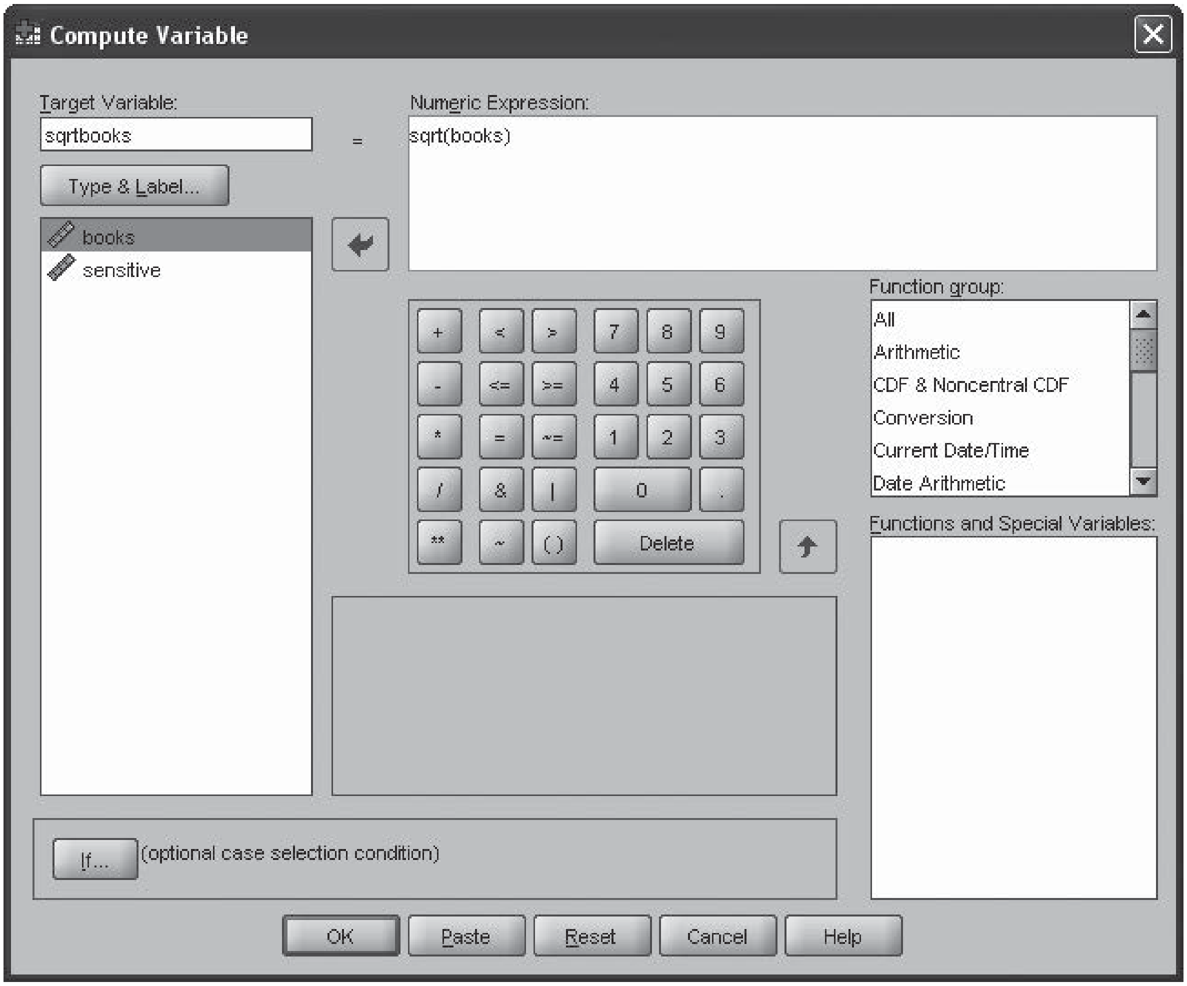
❷ 🖱 *Transform.*

❸ 🖱 *Compute Variable.* This will bring up a "Compute Variable" window.

❹ Name the new variable (which will be the square root of the scores for the "books" variable) by typing "sqrtbooks" in the "Target Variable" box. (You could give any name to the new variable, but it is best to give it a name that describes how it was figured, and we recommend keeping your SPSS variable names to 10–12 characters or less.)

❺ Type "sqrt(books)" in the "Numeric Expression" box. This tells SPSS to take the square root of each score for the "books" variable and create a new variable with those transformed scores. The Compute Variable window should now look like Figure 14.

❻ 🖱 *OK.*

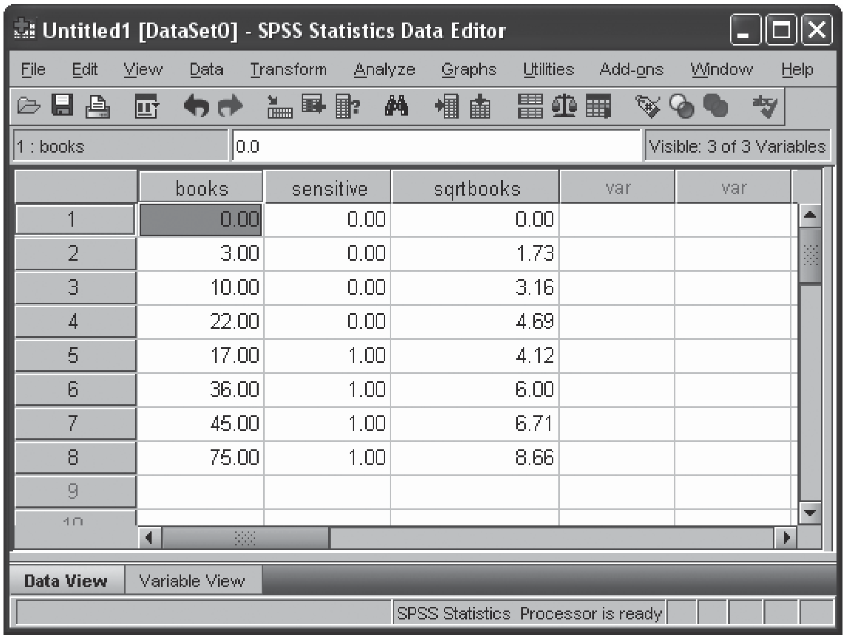


Compute Variable

Target Variable: sqrtbooks

Numeric Expression: sqrt(books)

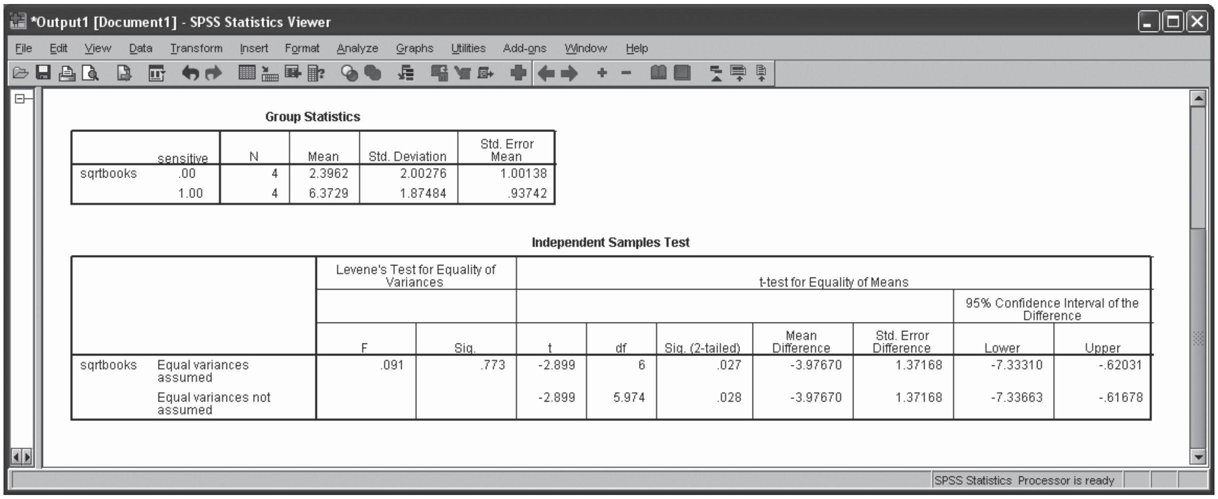
Figure 14 SPSS compute variable window for figuring the square root of the books scores for the fictional study comparing highly sensitive and not highly sensitive children on the number of books read in the past year.


Untitled1 [DataSet0] - SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
1 : books
0.0
Visible: 3 of 3 Variables
Data View
Variable View
SPSS Statistics Processor is ready

	books	sensitive	sqrtbooks	var	var
1	0.00	0.00	0.00		
2	3.00	0.00	1.73		
3	10.00	0.00	3.16		
4	22.00	0.00	4.69		
5	17.00	1.00	4.12		
6	36.00	1.00	6.00		
7	45.00	1.00	6.71		
8	75.00	1.00	8.66		
9					
10					

Figure 15 SPSS data editor window for the fictional study comparing highly sensitive and not highly sensitive children on the number of books read in the past year, including the scores for the "books" variable after a square root transformation.

Your SPSS data editor window should now look like Figure 15. You can now use the "sqrtbooks" scores in a *t* test for independent means to compare the scores of not highly sensitive and highly sensitive children. In Figure 16 we show the SPSS output of such a *t* test for independent means. The results of this *t* test are the same (within rounding error) as the results in Table 12.


*Output1 [Document1] - SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Add-ons Window Help
SPSS Statistics Processor is ready

Group Statistics

	sensitive	N	Mean	Std. Deviation	Std. Error Mean
sqrtbooks	.00	4	2.3962	2.00276	1.00138
	1.00	4	6.3729	1.87484	.93742

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
sqrtbooks	Equal variances assumed	.091	.773	-2.899	6	.027	-3.97670	1.37168	-7.33310	-.62031
	Equal variances not assumed			-2.899	5.974	.028	-3.97670	1.37168	-7.33663	-.61678

Figure 16 SPSS output window for a *t* test for independent means for the square-root transformed book scores from the fictional study comparing highly sensitive and not highly sensitive children on the number of books read in the past year.

Rank-Order Tests

We will use the example of the Wilcoxon rank-sum test for the scores from the study comparing the highly sensitive and not highly sensitive children on the number of books read in the past year (see Table 14).

❶ Enter the scores into SPSS. As shown in Figure 13, the score for each child is shown in the "books" column. The scores in the "sensitive" column show whether the child was not highly sensitive (a score of "0") or highly sensitive (a score of "1").

❷ 🖱 *Analyze.*

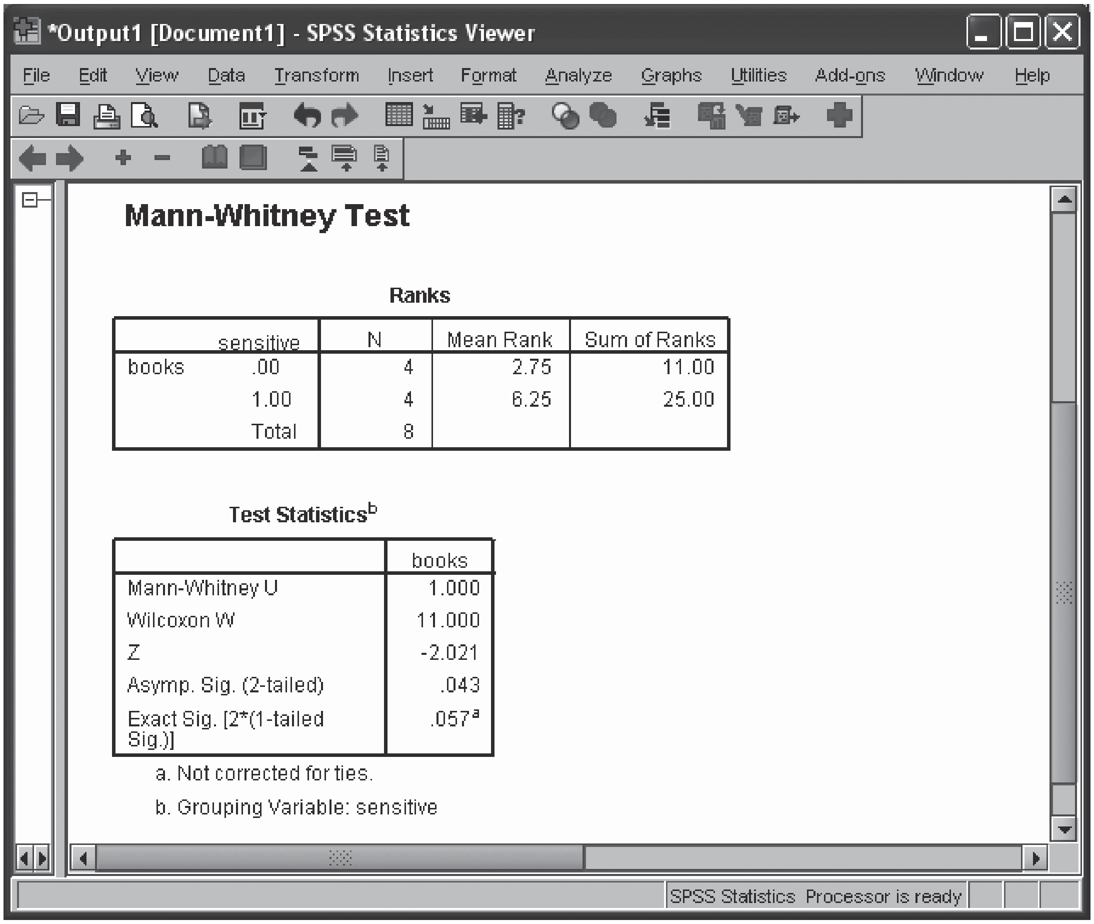
❸ 🖱 *Nonparametric tests.*

❹ 🖱 *2 Independent Samples.*

❺ 🖱 on the variable called "books" and then 🖱 the arrow next to the box labeled "Test Variable List." This tells SPSS that the rank-order test should be carried out on the scores for the "books" variable.

❻ 🖱 the variable called "sensitive" and then 🖱 the arrow next to the box labeled "Grouping Variable." This tells SPSS that the variable called "sensitive" shows which person is in which group. 🖱 *Define Groups.* You now tell SPSS the values you used to label each group. Put "0" in the Group 1 box and put "1" in the Group 2 box. 🖱 *Continue.*

❼ 🖱 *OK.* Your SPSS output window should look like Figure 17.


*Output1 [Document1] - SPSS Statistics Viewer
File
Edit
View
Data
Transform
Insert
Format
Analyze
Graphs
Utilities
Add-ons
Window
Help
SPSS Statistics Processor is ready

Mann-Whitney Test

Ranks

	sensitive	N	Mean Rank	Sum of Ranks
books	.00	4	2.75	11.00
	1.00	4	6.25	25.00
	Total	8		

Test Statistics[b]

	books
Mann-Whitney U	1.000
Wilcoxon W	11.000
Z	-2.021
Asymp. Sig. (2-tailed)	.043
Exact Sig. [2*(1-tailed Sig.)]	.057[a]

a. Not corrected for ties.

b. Grouping Variable: sensitive

Figure 17 SPSS output window for a Wilcoxon rank-sum test for the fictional study comparing highly sensitive and not highly sensitive children on the number of books read in the past year.

You will notice that the output gives the heading of “Mann-Whitney Test.” The Mann-Whitney test and the Wilcoxon rank-sum test differ in their computations but give mathematically equivalent final results. (As you will see in the second table in the SPSS output, the significance level is the same for Mann-Whitney and Wilcoxon rank-sum tests.) The first table in the SPSS output (labeled “Ranks”) provides information about the two variables. The first column gives the levels of the “sensitive” grouping variable (0 and 1, which indicate the not highly sensitive and highly sensitive groups, respectively.) The second, third, and fourth columns give, respectively, the number of individuals (*N*), mean rank, and sum of ranks for each group.

The second table in the SPSS output (labeled “Test Statistics”) shows the actual results of the nonparametric rank-ordered tests. We focus here on the results for the Wilcoxon rank-sum test (but the overall significance level and conclusion is the same, regardless of which test result you consider). Notice that the value of 11.000 is the same as the sum of the ranks for the not highly sensitive group, as shown in Table 14. The exact significance level of .043 for this result (shown in the “Asymp. Sig. (2-tailed)” row) is for a two-tailed test. In this example, we were using a one-tailed test, so this two-tailed significance of .043 represents a one-tailed significance of .043/2, which is .0215. This significance level of .0215 is less than our .05 cutoff for this example. This means that you can reject the null hypothesis and the research hypothesis is supported.

1. (a) χ^2 needed $(df = 5 - 1 = 4; p < .05) = 9.488$.

Category	*O*	Expected	*O* − *E*	(*O* − *E*)^2	(*O* − *E*)^2/*E*
A	19	(.2)(50) = 10	9	81	8.10
B	11	(.2)(50) = 10	1	1	.10
C	10	(.4)(50) = 20	−10	100	5.00
D	5	(.1)(50) = 5	0	0	0.00
E	5	(.1)(50) = 5	0	0	0.00
Total	50	50	0		$\chi^2 = 13.20$
Conclusion: Reject the null hypothesis					

(b) χ^2 needed $= 5.992$; $\chi^2 = 44.45$, reject the null hypothesis; (c) χ^2 needed $= 7.815$; $\chi^2 = 1.23$, do not reject the null hypothesis.

2. (a)

❶ **Restate the question as a research hypothesis and a null hypothesis about the populations.** There are two populations of interest:

Population 1: Clients like those of this social service agency.
Population 2: Clients for whom season makes no difference in when they use the social service agency.

The research hypothesis is that the distribution over seasons of when clients use the social service agency is different between the two populations. The null hypothesis is that the distributions over seasons of when clients use the social service agency is not different between the two populations.

❷ **Determine the characteristics of the comparison distribution.** Chi-square distribution with 3 degrees of freedom $(df = N_{\text{Categories}} - 1 = 4 - 1 = 3)$.

❸ **Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected.** .05 level, $df = 3$: χ^2 needed $= 7.815$.

❹ **Determine your sample’s score on the comparison distribution.**

Season	*O*	Expected	*O* − *E*	(*O* − *E*)^2	(*O* − *E*)^2/*E*
Winter	28	(1/4)(128) = 32	−4	16	.50
Spring	33	(1/4)(128) = 32	1	1	.03
Summer	16	(1/4)(128) = 32	−16	256	8.00
Fall	51	(1/4)(128) = 32	19	361	11.28
Total	128	128	0		$\chi^2 = 19.81$

❺ **Decide whether to reject the null hypothesis.** 19.81 is larger than 7.815; reject the null hypothesis; the research hypothesis is supported.

(b) In this study, the hypothesis testing involves a single nominal variable (that is, a variable with values that are categories). The appropriate statistical test in this situation is a chi-square test for goodness of fit. In this example, this will test whether the number of clients visiting the social service agency is equal across each of the four seasons of the year. The research hypothesis here is that the distribution over seasons of when clients use the social service agency is different for the two populations. (Population 1 is clients like those of this social service agency and Population 2 is clients for whom season makes no difference in when they use the social service agency.) The null hypothesis is that the distributions over seasons of when clients use the social service agency is not different between the two populations. If the season makes no difference, you would expect about 25% of clients to use the social service agency each season. (Last year there were 128 clients, so if season made no difference, we would expect 25% of 128, or 32 to use it each season.) Last year's actual numbers in each season (28, 33, 16, and 51) were clearly different from these expectations of 32 each season. But were they so different from these expectations that you should conclude that, in general, the numbers of new clients are not equally distributed over the seasons? The chi-square statistic reflects the discrepancy between observed (the actual numbers seen in the study) and the expected numbers. These numbers are called frequencies because they refer to how often or how frequently something happens (for example, an observed frequency of 28 clients last winter). For each category (such as the four seasons), you figure that observed minus expected discrepancy, square it, and divide by the expected frequency; then you add up the results. Chi-square uses squared discrepancies so that the result is not affected by the directions of the differences. You divide by the expected number to reduce the impact of the raw number of cases on the result. Applying this method to the present example, for the winter, 28 less than 32 is -4; -4 squared is 16; 16 divided by 32 is .50. Doing the same for the other three seasons, and then adding up the four, gives a total chi-square of 19.81.

Statisticians have determined mathematically what would happen if you took an infinite number of samples from a population with a fixed proportion of cases in each category and figured chi-square for each sample. This distribution depends only on how many categories are free to take on different expected values. (The total number expected is the total number of cases; thus, if you know the expected for any three categories, you can just subtract to get the number expected for the fourth.) A table of the chi-square distribution when three categories are free to vary shows that if the null hypothesis were true, there would be only a 5% chance of getting a chi-square of 7.815 or greater. Because our chi-square is larger than this, the observed result differs from the expected more than you would reasonably expect by chance. Thus, the number of new clients, in the long run, is probably not equal over the four seasons.

3. For (a), (b), and (c): $df = (N_{\text{Columns}} - 1)(N_{\text{Rows}} - 1) = (2 - 1)(2 - 1) = 1$; χ^2 needed $= 6.635$.
(a)

10 (13)	16 (13)	26 (50%)
16 (13)	10 (13)	26 (50%)
26	26	52

$\chi^2 = (10 - 13)^2/13 + (16 - 13)^2/13 + (16 - 13)^2/13 + (10 - 13)^2/13 = 2.77$. Do not reject the null hypothesis. Effect size, $\phi = \sqrt{\chi^2/N} = \sqrt{2.77/52} = \sqrt{.053} = .23$.
(b)$\chi^2 = .36$, do not reject the null hypothesis; Effect size, $\phi = .03$. (c) $\chi^2 = 27.68$, reject the null hypothesis. Effect size, $\phi = .23$. For (d), (e), and (f): $df = (N_{\text{Columns}} - 1)(N_{\text{Rows}} - 1) = (3 - 1)(2 - 1) = 2$; χ^2 needed $= 9.211$. (d) $\chi^2 = 2.76$, do not reject the null hypothesis; Effect size, Cramer's $\phi = \sqrt{\chi^2/[(N)(df_{\text{Smaller}})]} = \sqrt{2.76/[(72)(1)]} = \sqrt{.0383} = .20$. (e) $\chi^2 = 2.76$, do not reject the null hypothesis; Effect size, Cramer's $\phi = .18$. (f) $\chi^2 = 3.71$, do not reject the null hypothesis. Effect size, Cramer's $\phi = .22$.

4. (a)
❶ **Restate the question as a research hypothesis and a null hypothesis about the populations.** There are two populations of interest:

Population 1: People like those surveyed.
Population 2: People for whom the community they live in is independent of their opinion on the upcoming ballot initiative.

The research hypothesis is that the two populations are different (the community in which people live is not independent of their opinions on the upcoming ballot initiative). The null hypothesis is that the two populations are the same (the community in which people live is independent of their opinions on the upcoming ballot initiative).

❷ **Determine the characteristics of the comparison distribution.** Chi-square distribution with four degrees of freedom. $df = (N_{\text{Columns}} - 1)(N_{\text{Rows}} - 1) = (3 - 1)(3 - 1) = 4$

❸ **Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected.** .05 level, $df = 4$: χ^2 needed $= 9.488$.

❹ **Determine your sample's score on the comparison distribution.**

	Community			
	A	B	C	Total
For	12 (9.8)	6 (4.2)	3 (7)	21 (23.33%)
Against	18 (16.8)	3 (7.2)	15 (12)	36 (40.00%)
No opinion	12 (15.4)	9 (6.6)	12 (11)	33 (36.67%)
Total	42	18	30	90

$\chi^2 = (12 - 9.8)^2/9.8 + (6 - 4.2)^2/4.2 + (3 - 7)^2/7 + (18 - 16.8)^2/16.8 + (3 - 7.2)^2/7.2 + (15 - 12)^2/12 + (12 - 15.4)^2/15.4 + (9 - 6.6)^2/6.6 + (12 - 11)^2/11 = .49 + .77 + 2.29 + .09 + 2.45 + .75 + .75 + .87 + .09 = 8.55$.

❺ **Decide whether to reject the null hypothesis.** χ^2 in Step ❹ (8.55) is less extreme than Step ❸ cutoff (9.488). Therefore, do not reject the null hypothesis; the study is inconclusive.

(b) Cramer's $\phi = \sqrt{[8.55/(90)(2)]} = \sqrt{[8.55/180]} = \sqrt{.05} = .22$.

From Table 8, approximate power for a medium effect size and total $N = 100$ (the closest to the $N = 90$ in the study) is .66.

(c) In this study, we are testing a hypothesis with two nominal variables (that is, variables with values that are categories). The appropriate statistical test in this situation is a chi-square test for independence. This test will examine whether the community in which people live is related to their opinions on the upcoming ballot initiative. The research hypothesis in this situation is that the community in which people live is not independent of their opinions about the ballot initiative. The null hypothesis is that the community in which people live is independent of their opinions about the ballet initiative. In this example, 23.33% of all survey respondents were for the ballot initiative. Thus, if community is not related to opinion, 23.33% of the people in each community should be for the initiative. For example, you'd expect 9.8 of the 42 people surveyed in Community C to be for the initiative. Are the survey results so discrepant from these expectations that we should conclude that the community in which people live is related to their opinion on the upcoming ballot initiative?

Chi-square is a measure of the degree of difference between observed and expected results. For each combination of the 3 × 3 arrangement, you figure that discrepancy between observed and expected, square it, and divide by the expected number; then you add up the results. In the For–Community A combination, 12 minus 9.8 is 2.2, squared is 4.84, divided by 9.8 is .49 (rounded off). Doing the same for the other eight combinations and adding them all up gives 8.55.

Chi-square uses squared discrepancies so that the result is not affected by the directions of the differences. It is divided by the expected number to adjust for the relatively different numbers expected in the combinations.

Statisticians have determined mathematically what would happen if you took an infinite number of samples from a population with a fixed proportion of people in each of several groupings and figured the chi-square for each such sample. The distribution of such chi-squares depends only on the number of groupings free to take on different expected values. We always presume you know the row and column totals. Thus, for each community, if you know the numbers For and Against, you can figure out how many have no opinion. Furthermore, you need to know only two of the communities—say A and B—for any particular opinion, and you can figure out the third by subtracting these from the total of that row. So only four combinations—say For and Against for Communities A and B—are "free to vary."

A table of the chi-square distribution when four groupings are free to vary shows that there is only a 5% chance of getting a chi-square of 9.488 or greater if the null hypothesis is true. Because our chi-square is smaller than this, the observed numbers in each category differ from the expected numbers less than they would need to before we would be willing to reject the null hypothesis that a person's opinion is unrelated to his or her community. The survey is inconclusive.

We can, however, estimate the actual degree of linkage in this group surveyed between community and opinion. The procedure is called "Cramer's phi," figured by dividing your chi-square by the number of people included in the analysis times the degrees of freedom of the smaller side of the table, then taking the square root of the results. In this example, this comes out to .22.

This statistic ranges from 0 (no relationship) to 1 (a perfect relationship—knowing a person's status on one of the dimensions, such as what community they are from, would let you perfectly predict their status on the other dimension, such as their opinion). Thus, .22 is a fairly low figure, although given the amount of other things that affect any relationship, by the standards of behavorial and social science research, a Cramer's phi of .22 would be considered a medium-sized relation. (To be exact, a Cramer's phi of .21 is the number given for a medium effect size.)

Looking at this another way, we can ask, if there really is a moderate relationship between opinion and community in the population, what is the chance that this whole process would have led to a positive conclusion? Statisticians have provided tables that give this probability. In this situation, it turns out that there would be about a 66% chance. If there were truly a large effect in the population (which would be a Cramer's phi of about .35), there is a 99% chance we would have come to a positive conclusion. Thus, given the result of this study, if any relationship exists, it is almost surely not a large one.

5. (a) $\phi = \sqrt{\chi^2/N} = \sqrt{16/100} = .40$;
(b) Cramer's $\phi = \sqrt{\chi^2/[(N)(df_{\text{Smaller}})]} = \sqrt{16/[(100)(1)]}$ $= .40$;
(c) Cramer's $\phi = .28$;
(d) $\phi = .28$;
(e) $\phi = .28$.
6. From Table 8: (a) .08; (b) .32; (c) .11; (d) .07; (e) .06; (f) .06.
7. From Table 9: (a) 87; (b) 26; (c) 133; (d) 133; (e) 39.
8. (a) You should get the same results within rounding error. (b) $\phi = \sqrt{\chi^2/N} = \sqrt{5.55/69} = .28$; (c) Similar to 4c but focusing on this study's results.
9. (a) 4, 2, 3, 5, 6; (b) 5.92, 3.78, 3.61, 3.59, 4.24.
10. (a) t needed $(df = 8, p < .05, \text{two-tailed}) = -2.306$, 2.306; Group A: $M = 3.8$, $S^2 = 5.06$; Group B: $M = 5.7$, $S^2 = 6.76$; $S^2_{\text{Pooled}} = 5.91$; $S_{\text{Difference}} = 1.54$; $t = -1.23$; do not reject the null hypothesis. (b) Group A: 1.1, 1.6, 2.1, 1.9, 2.7; Group B: 1.4, 3.0, 2.4, 2.6, 2.2. (c) t needed $= -2.306$, 2.306; Group A: $M = 1.88$, $S^2 = .35$; Group B: $M = 2.32$, $S^2 = .35$; $S^2_{\text{Pooled}} = .35$; $S_{\text{Difference}}$ $= .37$; $t = -1.19$; do not reject the null hypothesis. (d) It would not have been correct to carry out a t test on the numbers as they were (without transforming them). This is because the distributions of the samples were very skewed for both language groups. Thus, it seemed likely that the population distributions were also seriously skewed. That would clearly violate the assumption for a t test that the underlying population distributions are normal. Thus, I took the square root of each score. This had the advantage of making the sample distributions much closer to normal. This suggests that the population distributions of square roots of family sizes are probably nearly normally distributed. I realize that taking the square root of each family size distorts its straightforward meaning. However, the impact for the individuals in the family of each additional child is probably not equal. That is, going from no children to one child has a huge impact. Going from one to two has less, and going from seven to eight probably makes much less difference for the family.

In any case, having taken the square root of each score, I then carried out an ordinary *t* test for independent means using these square-root transformed scores. As with the original *t* test, the result was inconclusive (the null hypothesis could not be rejected)—but at least I could be confident that I had done the analysis correctly.

11. (a) *F* needed $(df = 2, 6; p < .01) = 10.93$; Sad: $M = 446$, $S^2 = 47{,}089$; Angry: $M = 259$, $S^2 = 11{,}727$; Exuberant: $M = 918.67$, $S^2 = 7{,}184$; $S^2_{\text{Between}} = 346{,}775.22$; $S^2_{\text{Within}} = 22{,}000$; $F = 15.76$; reject the null hypothesis. (b) 14.2, 22.9, 24.8; 11.7, 18.4, 17.3; 28.9, 30.2, 31.7. (c) $M = 20.63$, $S^2 = 31.94$; $M = 15.8$, $S^2 = 12.91$; $M = 30.27$, $S^2 = 1.96$; $M = 30.27$, $S^2 = 1.96$; $S^2_{\text{Between}} = 162.82$; $S^2_{\text{Within}} = 15.60$; $F = 10.44$; do not reject the null hypothesis. (d) Similar to 10d above, except note that the square root transformation does *not* solve the problem of skew and that it also creates distributions very likely to violate the assumption of equal population variances.
12. Miller wanted to examine the relationships among the variables he was studying, probably including various parametric hypothesis-testing techniques such as the *t* test or an analysis of variance. Such procedures are based on the assumption that the distributions of the variables in the population follow a normal curve. However, Miller first checked the distributions of the variables he was studying and found that the scores on two key measures were skewed, suggesting that the population distributions for these variables probably violated the normal distribution assumption. (The rest of your answer should be similar to 10d above.)

Steps of Hypothesis Testing for Major Procedures

Chi-square test for goodness of fit

① **Restate the question as a research hypothesis and a null hypothesis about the populations.**

② **Determine the characteristics of the comparison distribution.** Chi-square distribution, $df = N_{\text{Categories}} - 1$.

③ **Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected.** Use chi-square table.

④ **Determine your sample's score on the comparison distribution.** $\chi^2 = \Sigma\,[(O - E)^2/E]$

Ⓐ Determine the actual, observed frequencies in each category.
Ⓑ Determine the expected frequencies in each category.
Ⓒ In each category, take observed minus expected frequencies.
Ⓓ Square each of these differences.
Ⓔ Divide each squared difference by the expected difference for its category.
Ⓕ Add up the results of Step Ⓔ for all the categories.

⑤ **Decide whether to reject the null hypothesis.** Compare scores from Steps ③ and ④.

Chi-square test for independence

① **Restate the question as a research hypothesis and a null hypothesis about the populations.**

② **Determine the characteristics of the comparison distribution.** Chi-square distribution, $df = (N_{\text{Columns}} - 1)\,(N_{\text{Rows}} - 1)$.

③ **Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected.** Use chi-square table.

④ **Determine your sample's score on the comparison distribution.** $E = (R/N)(C)$; $\chi^2 = \Sigma[(O - E)^2/E]$

Ⓐ Determine the actual, observed frequencies in each cell.
Ⓑ Determine the expected frequencies in each cell:
(i) Find each row's percentage of the total.
(ii) For each cell, multiply its row's percentage by its column's total
Ⓒ In each cell, take observed minus expected frequencies.
Ⓓ Square each of these differences.
Ⓔ Divide each squared difference by the expected frequency for its cell.
Ⓕ Add up the results of Step Ⓔ for all the cells.

⑤ **Decide whether to reject the null hypothesis.** Compare scores from Steps ③ and ④.