Chapter 11

Inference for Distributions of Categorical Data

case study

Do Dogs Resemble Their Owners?

Some people look a lot like their pets. Maybe they deliberately choose animals that match their appearance. Or maybe we're perceiving similarities that aren't really there. Researchers at the University of California, San Diego, decided to investigate. They designed an experiment to test whether or not dogs resemble their owners. The researchers believed that resemblance between dog and owner might differ for purebred and mixed-breed dogs.

A random sample of 45 dogs and their owners was photographed separately at three dog parks. Then, researchers "constructed triads of pictures, each consisting of one owner, that owner's dog, and one other dog photographed at the same park." The subjects in the experiment were 28 undergraduate psychology students. Each subject was presented with the individual sets of photographs and asked to identify which dog belonged to the pictured owner. A dog was classified as resembling its owner if more than half of the 28 undergraduate students matched dog to owner.[1]

The table below summarizes the results. There is some support for the researchers' belief that resemblance between dog and owner might differ for purebred and mixed-breed dogs.

	Breed status	
Resemblance?	**Purebred dogs**	**Mixed-breed dogs**
Resemble owner	16	7
Don't resemble	9	13

Do these data provide convincing evidence of an association between dogs' breed status and whether or not they resemble their owners? By the end of this chapter, you will have developed the tools you need to answer this question.

Introduction

In the previous chapter, we discussed inference procedures for comparing the proportion of successes for two populations or treatments. Sometimes we want to examine the distribution of a single categorical variable in a population. The *chi-square test for goodness of fit* allows us to determine whether a hypothesized distribution seems valid. This method is useful in a field like genetics, where the laws of probability give the expected proportion of outcomes in each category.

We can decide whether the distribution of a categorical variable differs for two or more populations or treatments using a *chi-square test for homogeneity*. In doing so, we will often organize our data in a two-way table. It is also possible to use the information in a two-way table to study the relationship between two categorical variables. The *chi-square test for independence* allows us to determine if there is convincing evidence of an association between the variables in the population at large.

The methods of this chapter help us answer questions such as these:

- Are the birthdays of NHL players evenly distributed throughout the year?
- Does background music influence customer purchases?
- Is there an association between anger level and heart disease?

Of course, we have to do a careful job of describing the data before we proceed to statistical inference. In Chapter 1, we discussed graphical and numerical methods of data analysis for categorical variables. You may want to quickly review Section 1.1 now.

Here's an Activity that gives you a taste (pardon the pun) of what lies ahead.

ACTIVITY The Candy Man Can

MATERIALS:

Large bag of M&M'S® Milk Chocolate Candies for the class; TI-83/84 or TI-89 for each team of 3 to 4 students

Mars, Incorporated, which is headquartered in McLean, Virginia, makes milk chocolate candies. Here's what the company's Consumer Affairs Department says about the color distribution of its M&M'S Milk Chocolate Candies:

> *On average, M&M'S Milk Chocolate Candies will contain 13 percent of each of browns and reds, 14 percent yellows, 16 percent greens, 20 percent oranges and 24 percent blues.*

The purpose of this activity is to determine whether the company's claim is believable.

1. Your teacher will take a random sample of 60 M&M'S from a large bag and give one or more pieces of candy to each student. As a class, count the number of M&M'S® Chocolate Candies of each color. Make a table on the board that summarizes these *observed counts*.

2. How can you tell if the sample data give convincing evidence to dispute the company's claim? Each team of three or four students should discuss this question and devise a formula for a test statistic that measures the difference between the observed and expected color distributions. The test statistic should yield a

single number when the observed and expected values are plugged in. Here are some questions for your team to consider:

- Should we look at the difference between the observed and expected *proportions* in each color category or between the observed and expected *counts* in each category?
- Should we use the differences themselves, the absolute value of the differences, or the square of the differences?
- Should we divide each difference value by the sample size, expected count, or nothing at all?

3. Each team will share its proposed test statistic with the class. Your teacher will then reveal how the *chi-square statistic* χ^2 is calculated.

4. Discuss as a class: If your sample is consistent with the company's claimed distribution of M&M'S® Chocolate Candies colors, will the value of χ^2 be large or small? If your sample is not consistent with the company's claimed color distribution, will the value of χ^2 be large or small?

5. Compute the value of the chi-square test statistic for the class's data. Is this value large or small? To find out, you and your classmates will perform a simulation.

6. Suppose that the company's claimed color distribution is correct. We'll use numerical labels from 1 to 100 to represent the color of a randomly chosen M&M'S Milk Chocolate Candy:

1–13 = brown 14–26 = red 27–40 = yellow 41–56 = green
57–76 = orange 77–100 = blue

Use the calculator command below to simulate choosing a random sample of 60 candies.

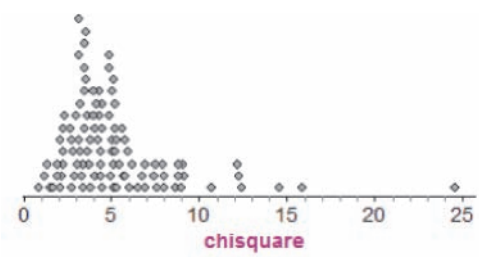
TI-83/84: `RandInt(1,100,60)` → `L1`
TI-89: `tistat.randint(1,100,60)` → `list1`

Sort the list in ascending order. Then record the observed counts in each color category and compute the value of χ^2 for your simulated sample.

7. Your teacher will draw and label axes for a class dotplot. Plot the result you got in Step 6 on the graph.

8. Repeat Steps 6 and 7 if needed to get a total of at least 40 repetitions of the simulation for your class.

9. Based on the class's simulation results, how surprising would it be to get a χ^2-value as large as or larger than the one you did in Step 5 by chance alone when sampling from the claimed distribution? What conclusion would you draw?



Here is an example of what the class dotplot in the Activity might look like after 100 trials. The graph shows what values of the chi-square statistic are likely to occur by chance alone when sampling from the company's claimed M&M'S® Chocolate Candies color distribution. Where did your class's χ^2-value fall? You will learn more about the sampling distribution of the chi-square statistic shortly.

11.1 Chi-Square Tests for Goodness of Fit

WHAT YOU WILL LEARN By the end of the section, you should be able to:

- State appropriate hypotheses and compute expected counts for a chi-square test for goodness of fit.
- Calculate the chi-square statistic, degrees of freedom, and *P*-value for a chi-square test for goodness of fit.
- Perform a chi-square test for goodness of fit.
- Conduct a follow-up analysis when the results of a chi-square test are statistically significant.

Jerome's class did the Candy Man Can Activity. The **one-way table** below summarizes the data from the class's sample of M&M'S® Milk Chocolate Candies. In general, one-way tables display the distribution of a single categorical variable for the individuals in a sample.

Color:	Blue	Orange	Green	Yellow	Red	Brown	**Total**
Count:	9	8	12	15	10	6	**60**

The sample proportion of blue M&M'S is $\hat{p} = \frac{9}{60} = 0.15$. Because the company claims that 24% of all M&M'S Milk Chocolate Candies are blue, Jerome might believe that something fishy is going on. We could use the one-sample z test for a proportion from Chapter 9 to test the hypotheses

$$H_0: p = 0.24$$
$$H_a: p \neq 0.24$$

where p is the true population proportion of blue M&M'S® Chocolate Candies. We could then perform additional significance tests for each of the remaining colors.

Note that the correct alternative hypothesis H_a is two-sided. A sample proportion of blue M&M'S much higher or much lower than 0.24 would give Jerome reason to be suspicious about the company's claim. It's not appropriate to adjust H_a after looking at the sample data!

Not only would this method be fairly inefficient, it would also lead to the problem of multiple comparisons, which we'll discuss in Section 11.2. More important, this approach wouldn't tell us how likely it is to get a random sample of 60 candies with a color *distribution* that differs as much from the one claimed by the company as the class's sample does, taking all the colors into consideration at one time. For that, we need a new kind of significance test, called a **chi-square test for goodness of fit**.

Comparing Observed and Expected Counts: The Chi-Square Statistic

As with any test, we begin by stating hypotheses. The null hypothesis in a chi-square test for goodness of fit should state a claim about the distribution of a single categorical variable in the population of interest. In the case of the Candy Man Can Activity, the categorical variable we're measuring is color and the

population of interest is all M&M'S® Milk Chocolate Candies. The appropriate null hypothesis is

H_0: The company's stated color distribution for all M&M'S Milk Chocolate Candies is correct.

The alternative hypothesis in a chi-square test for goodness of fit is that the categorical variable does not have the specified distribution. For the M&M'S, our alternative hypothesis is

H_a: The company's stated color distribution for all M&M'S Milk Chocolate Candies is *not* correct.

THINK ABOUT IT

Why did we state hypotheses in words for a chi-square test for goodness of fit? We can also write the hypotheses in symbols as

$$H_0: p_{\text{blue}} = 0.24,\ p_{\text{orange}} = 0.20,\ p_{\text{green}} = 0.16,$$
$$p_{\text{yellow}} = 0.14,\ p_{\text{red}} = 0.13,\ p_{\text{brown}} = 0.13,$$

H_a: At least two of the p_i's are incorrect

where p_{color} = the true population proportion of M&M'S Milk Chocolate Candies of that color. Why don't we write the alternative hypothesis as H_a: At least one of the p_i's is incorrect? If the stated proportion in one category is wrong, then the stated proportion in at least one other category must be wrong because the sum of the p_i's must be 1.

Don't state the alternative hypothesis in a way that suggests that *all* the proportions in the hypothesized distribution are wrong. For instance, it would be *incorrect* to write

$$H_a: p_{\text{blue}} \neq 0.24,\ p_{\text{orange}} \neq 0.20,\ p_{\text{green}} \neq 0.16,$$
$$p_{\text{yellow}} \neq 0.14,\ p_{\text{red}} \neq 0.13,\ p_{\text{brown}} \neq 0.13$$

The idea of the chi-square test for goodness of fit is this: we compare the **observed counts** from our sample with the counts that would be expected if H_0 is true. (*Remember*: we always assume that H_0 is true when performing a significance test.) The more the observed counts differ from the **expected counts,** the more evidence we have against the null hypothesis. In general, the expected counts can be obtained by multiplying the sample size by the proportion in each category according to the null hypothesis. Here's an example that illustrates the process.

EXAMPLE

Return of the M&M'S® Chocolate Candies

Computing expected counts

PROBLEM: Jerome's class collected data from a random sample of 60 M&M'S Milk Chocolate Candies. Calculate the expected counts for each color. Show your work.

SOLUTION: Assuming that the color distribution stated by Mars, Inc., is true, 24% of all M&M'S Milk Chocolate Candies produced are blue. For random samples of 60 candies, the average number of blue M&M'S should be (60)(0.24) = 14.40. This is our expected count of blue M&M'S® Chocolate Candies. Using this same method, we find the expected counts for the other color categories:

Orange: $(60)(0.20) = 12.00$

Green: $(60)(0.16) = 9.60$

Yellow: $(60)(0.14) = 8.40$

Red: $(60)(0.13) = 7.80$

Brown: $(60)(0.13) = 7.80$

Color	Observed	Expected
Blue	9	14.40
Orange	8	12.00
Green	12	9.60
Yellow	15	8.40
Red	10	7.80
Brown	6	7.80

For Practice *Try Exercise* 1

Did you notice that the expected count sounds a lot like the expected value of a random variable from Chapter 6? That's no coincidence. The number of M&M'S® Chocolate Candies of a specific color in a random sample of 60 candies is a binomial random variable. Its expected value is np, the average number of candies of this color in many samples of 60 M&M'S Milk Chocolate Candies. The expected value is not likely to be a whole number.

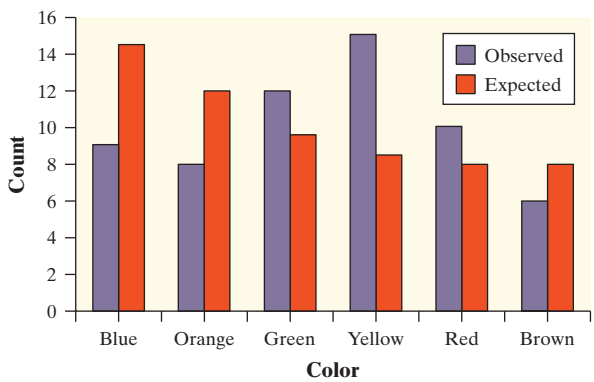


FIGURE 11.1 Bar graph comparing observed and expected counts for Jerome's class sample of 60 M&M'S® Milk Chocolate Candies.

To see if the data give convincing evidence for the alternative hypothesis, we compare the observed counts from our sample with the expected counts. If the observed counts are far from the expected counts, that's the evidence we were seeking. The table in the example gives the observed and expected counts for the sample of 60 M&M'S in Jerome's class. Figure 11.1 shows a side-by-side bar graph comparing the observed and expected counts.

We see some fairly large differences between the observed and expected counts in several color categories. How likely is it that differences this large or larger would occur just by chance in random samples of size 60 from the population distribution claimed by Mars, Inc.? To answer this question, we calculate a statistic that measures how far apart the observed and expected counts are. The statistic we use to make the comparison is the **chi-square statistic**

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

(The symbol χ is the lowercase Greek letter chi, pronounced "kye.")

DEFINITION: Chi-square statistic

The **chi-square statistic** is a measure of how far the observed counts are from the expected counts. The formula for the statistic is

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

where the sum is over all possible values of the categorical variable.

Let's use this formula to compare the observed and expected counts for Jerome's class sample.

EXAMPLE Return of the M&M'S® Chocolate Candies

Calculating the chi-square statistic

The table shows the observed and expected counts for the random sample of 60 M&M'S Milk Chocolate Candies in Jerome's class.

Color	Observed	Expected
Blue	9	14.40
Orange	8	12.00
Green	12	9.60
Yellow	15	8.40
Red	10	7.80
Brown	6	7.80

PROBLEM: Calculate the chi-square statistic.

SOLUTION: The formula for the chi-square statistic is

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

For Jerome's data, we add six terms—one for each color category:

$$\chi^2 = \frac{(9 - 14.40)^2}{14.40} + \frac{(8 - 12.00)^2}{12.00} + \frac{(12 - 9.60)^2}{9.60}$$
$$+ \frac{(15 - 8.40)^2}{8.40} + \frac{(10 - 7.80)^2}{7.80} + \frac{(6 - 7.80)^2}{7.80}$$
$$= 2.025 + 1.333 + 0.600 + 5.186 + 0.621 + 0.415 = 10.180$$

For Practice *Try Exercise* 3

THINK ABOUT IT

Why do we divide by the expected count when calculating the chi-square statistic? Suppose you obtain a random sample of 60 M&M'S Milk Chocolate Candies. Which would be more surprising: getting 18 blue candies or 12 yellow candies in the sample? Earlier, we computed the expected counts for these two categories as 14.4 and 8.4, respectively. The difference in the observed and expected counts for the two colors would be

Blue: $18 - 14.4 = 3.6$ Yellow: $12 - 8.4 = 3.6$

In both cases, the number of M&M'S® Chocolate Candies in the sample exceeds the expected count by the same amount. But it's much more surprising to be off by 3.6 out of an expected 8.4 yellow candies (almost a 50% discrepancy) than to be off by 3.6 out of an expected 14.4 blue candies (a 25% discrepancy). For that reason, we want the category with a larger *relative* difference to contribute more heavily to the evidence against H_0 and in favor of H_a measured by the χ^2 statistic.

If we just computed $(\text{Observed} - \text{Expected})^2$ for each category instead, the contributions of these two color categories would be the same:

Blue: $(18 - 14.40)^2 = 12.96$ Yellow: $(12 - 8.40)^2 = 12.96$

By using $\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$, we guarantee that the color category with the larger relative difference will contribute more heavily to the total:

$$\text{Blue: } \frac{(18 - 14.40)^2}{14.40} = 0.90 \quad \text{Yellow: } \frac{(12 - 8.40)^2}{8.40} = 1.54$$

Think of χ^2 as a measure of the distance of the observed counts from the expected counts. Like any distance, it is always zero or positive, and it is zero only when the observed counts are exactly equal to the expected counts. Large values of χ^2 are stronger evidence for H_a because they say that the observed counts are far from what we would expect if H_0 were true. Small values of χ^2 suggest that the data are consistent with the null hypothesis. Is $\chi^2 = 10.180$ a large value? You know the drill: compare the observed value 10.180 against the sampling distribution that shows how χ^2 would vary in repeated random sampling if the null hypothesis were true.

CHECK YOUR UNDERSTANDING

Mars, Inc., reports that their M&M'S® Peanut Chocolate Candies are produced according to the following color distribution: 23% each of blue and orange, 15% each of green and yellow, and 12% each of red and brown. Joey bought a randomly selected bag of Peanut Chocolate Candies and counted the colors of the candies in his sample: 12 blue, 7 orange, 13 green, 4 yellow, 8 red, and 2 brown.

1. State appropriate hypotheses for testing the company's claim about the color distribution of M&M'S Peanut Chocolate Candies.

2. Calculate the expected count for each color, assuming that the company's claim is true. Show your work.

3. Calculate the chi-square statistic for Joey's sample. Show your work.

The Chi-Square Distributions and *P*-Values

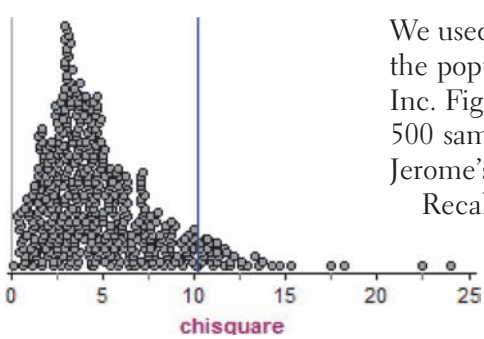


FIGURE 11.2 Fathom dotplot showing values of the chi-square statistic in 500 simulated samples of size $n = 60$ from the population distribution of M&M'S® Milk Chocolate Candies stated by the company.

We used Fathom software to simulate taking 500 random samples of size 60 from the population distribution of M&M'S Milk Chocolate Candies given by Mars, Inc. Figure 11.2 shows a dotplot of the values of the chi-square statistic for these 500 samples. The blue vertical line is plotted at the value of $\chi^2 = 10.180$ from Jerome's class data.

Recall that larger values of χ^2 give more convincing evidence against H_0 and in favor of H_a. According to Fathom, 37 of the 500 simulated samples resulted in a chi-square statistic of 10.180 or higher. Our estimated *P*-value is $37/500 = 0.074$. Because the *P*-value exceeds the default $\alpha = 0.05$ significance level, we fail to reject H_0. We do not have convincing evidence that the company's claimed color distribution is incorrect.

As Figure 11.2 suggests, the sampling distribution of the chi-square statistic is *not* a Normal distribution. It is a right-skewed distribution that allows only nonnegative values because χ^2 can never be negative.

The sampling distribution of χ^2 differs depending on the number of possible values for the categorical variable (that is, on the number of categories).

When the expected counts are all at least 5, the sampling distribution of the χ^2 statistic is modeled well by a **chi-square distribution** with degrees of freedom (df) equal to the number of categories minus 1. As with the t distributions, there is a different chi-square distribution for each possible value of df. Here are the facts.

THE CHI-SQUARE DISTRIBUTIONS

The chi-square distributions are a family of density curves that take only nonnegative values and are skewed to the right. A particular chi-square distribution is specified by giving its degrees of freedom. The chi-square test for goodness of fit uses the chi-square distribution with degrees of freedom = the number of categories − 1.

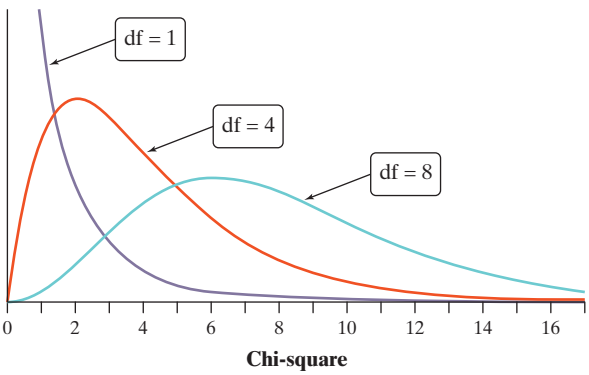


FIGURE 11.3 The density curves for three members of the chi-square family of distributions.

Figure 11.3 shows the density curves for three members of the chi-square family of distributions. As the degrees of freedom (df) increase, the density curves become less skewed, and larger values become more probable. Here are two other interesting facts about the chi-square distributions:

- The mean of a particular chi-square distribution is equal to its degrees of freedom.
- For df > 2, the mode (peak) of the chi-square density curve is at df − 2.

When df = 8, for example, the chi-square distribution has a mean of 8 and a mode of 6.

To get P-values from a chi-square distribution, we can use technology or Table C in the back of the book. The following example shows how to use the table.

EXAMPLE Return of the M&M'S® Chocolate Candies

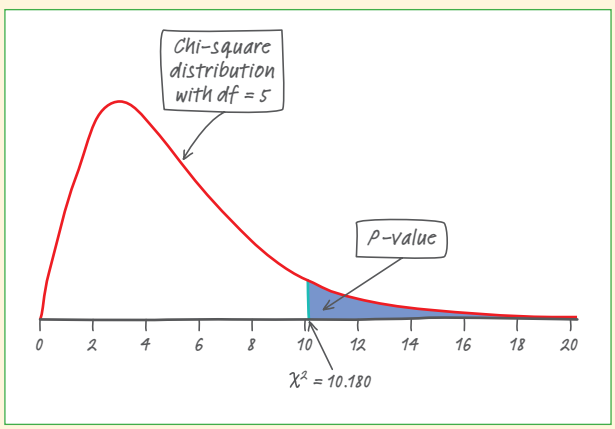


FIGURE 11.4 The P-value for a chi-square test for goodness of fit using Jerome's M&M'S® Chocolate Candies class data.

Finding the P-value

In the last example, we computed the chi-square statistic for the random sample of 60 M&M'S Milk Chocolate Candies in Jerome's class: $\chi^2 = 10.180$. Now let's find the P-value. Because all the expected counts are at least 5, the χ^2 statistic will be modeled well by a chi-square distribution when H_0 is true. There are 6 color categories for M&M'S Milk Chocolate Candies, so df = 6 − 1 = 5.

The P-value is the probability of getting a value of χ^2 as large as or larger than 10.180 when H_0 is true. Figure 11.4 shows this probability as an area under the chi-square density curve with 5 degrees of freedom.

To find the *P*-value using Table C, look in the df = 5 row. The value $\chi^2 = 10.180$ falls between the critical values 9.24 and 11.07. The corresponding areas in the right tail of the chi-square distribution with 5 degrees of freedom are 0.10 and 0.05. (See the excerpt from Table C on the right.) So the *P*-value for a test based on Jerome's data is between 0.05 and 0.10.

	P		
df	.15	.10	.05
4	6.74	7.78	9.49
5	8.12	9.24	11.07
6	9.45	10.64	12.59

Now let's look at how to find the *P*-value with your calculator.

25. TECHNOLOGY CORNER FINDING *P*-VALUES FOR CHI-SQUARE TESTS ON THE CALCULATOR

TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

To find the *P*-value in the M&M'S® example with your calculator, use the χ^2cdf(Chi-square Cdf on the TI-89) command. We ask for the area between $\chi^2 = 10.180$ and a very large number (we'll use 10,000) under the chi-square density curve with 5 degrees of freedom.

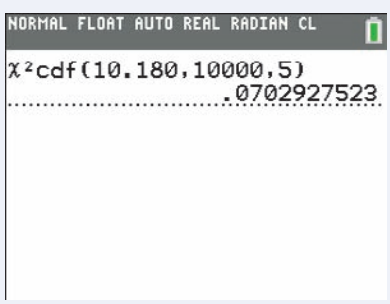
TI-83/84

- Press 2nd VARS (DISTR) and choose χ^2cdf(.

OS 2.55 or later: In the dialog box, enter these values: `lower:10.18, upper:10000, df:5`, choose `Paste`, and then press ENTER. **Older OS:** Complete the command χ^2`cdf(10.180,10000,5)` and press ENTER.

TI-89

- In the Stats/List Editor, Press F5 (Distr) and choose `Chi-square Cdf...`.
- In the dialog box, enter these values: `Lower value:10.18, Upper value:10000, Deg of Freedom, df:5`, and then choose ENTER.

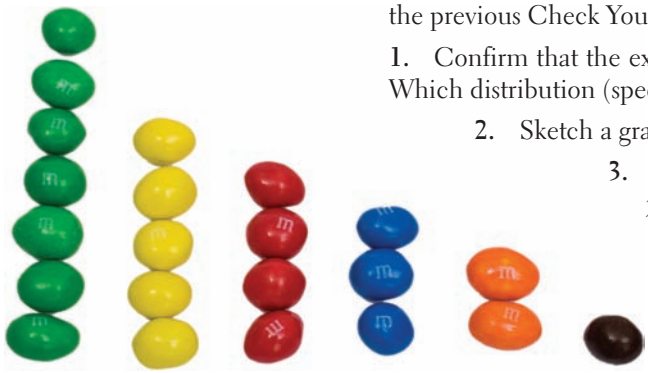


As the calculator screen shots show, this method gives a more precise *P*-value than Table C.

Table C gives us an interval in which the *P*-value falls. The calculator's χ^2`cdf` (`Chi-square Cdf` on the TI-89) command gives a result that is consistent with Table C but more precise. For that reason, we recommend using your calculator to compute *P*-values from a chi-square distribution.

Based on Jerome's sample, what conclusion can we draw about H_0: the company's stated color distribution for all M&M'S® Milk Chocolate Candies is correct? Because our *P*-value of 0.07 is greater than $\alpha = 0.05$, we fail to reject H_0. We don't have convincing evidence that the company's claimed color distribution is incorrect.

Failing to reject H_0 does not mean that the null hypothesis is true! That is, we can't conclude that the color distribution claimed by Mars, Inc., *is* correct. All we can say is that the sample data did not provide convincing evidence to reject H_0.



CHECK YOUR UNDERSTANDING

Let's continue our analysis of Joey's sample of M&M'S® Peanut Chocolate Candies from the previous Check Your Understanding (page 684).

1. Confirm that the expected counts are large enough to use a chi-square distribution. Which distribution (specify the degrees of freedom) should we use?
2. Sketch a graph like Figure 11.4 on page 685 that shows the *P*-value.
3. Use Table C to find the *P*-value. Then use your calculator's χ^2`cdf` command.
4. What conclusion would you draw about the company's claimed color distribution for M&M'S® Peanut Chocolate Candies? Justify your answer.

Carrying Out a Test

Like our test for a population proportion, the chi-square test for goodness of fit uses some approximations that become more accurate as we take larger samples. The **Large Counts condition** says that all expected counts must be at least 5. Before performing a test, we must also check that the Random and 10% conditions are met.

CONDITIONS FOR PERFORMING A CHI-SQUARE TEST FOR GOODNESS OF FIT

- **Random:** The data come from a well-designed random sample or randomized experiment.
  - **10%:** When sampling without replacement, check that $n \leq \frac{1}{10}N$.
- **Large Counts:** All *expected* counts are at least 5.

Before we start using the chi-square test for goodness of fit, we have two important cautions to offer.

1. The chi-square test statistic compares observed and expected *counts*. Don't try to perform calculations with the observed and expected *proportions* in each category.
2. When checking the Large Counts condition, be sure to examine the *expected* counts, not the observed counts.

We can also write these hypotheses symbolically using p_i to represent the proportion of individuals in the population that fall in category i:

H_0: $p_1 =$ ___, $p_2 =$ ___, …, $p_k =$ ___.

H_a: At least two of the p_i's are incorrect.

THE CHI-SQUARE TEST FOR GOODNESS OF FIT

Suppose the conditions are met. To determine whether a categorical variable has a specified distribution in the population of interest, expressed as the proportion of individuals falling into each possible category, perform a test of

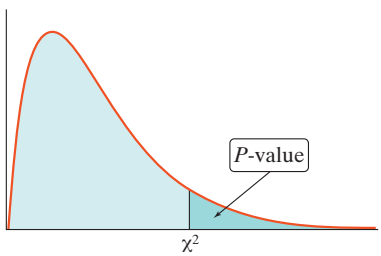
H_0: The stated distribution of the categorical variable in the population of interest is correct.

H_a: The stated distribution of the categorical variable in the population of interest is not correct.

Start by finding the expected count for each category assuming that H_0 is true. Then calculate the chi-square statistic

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

where the sum is over the k different categories. The P-value is the area to the right of χ^2 under the density curve of the chi-square distribution with $k - 1$ degrees of freedom.



The next example shows the chi-square test for goodness of fit in action. As always, we follow the four-step process when performing inference.

EXAMPLE Birthdays and Hockey

STEP 4

A test for equal proportions

In his book *Outliers*, Malcolm Gladwell suggests that a hockey player's birth month has a big influence on his chance to make it to the highest levels of the game. Specifically, because January 1 is the cut-off date for youth leagues in Canada (where many National Hockey League (NHL) players come from), players born in January will be competing against players up to 12 months younger. The older players tend to be bigger, stronger, and more coordinated and hence get more playing time, more coaching, and have a better chance of being successful.

To see if birth date is related to success (judged by whether a player makes it into the NHL), a random sample of 80 NHL players from a recent season was selected and their birthdays were recorded. The one-way table below summarizes the data on birthdays for these 80 players:

Birthday:	Jan–Mar	Apr–Jun	Jul–Sep	Oct–Dec
Number of players:	32	20	16	12

Do these data provide convincing evidence that the birthdays of NHL players are not uniformly distributed throughout the year?

The null hypothesis says that NHL players' birthdays are evenly distributed across the four quarters of the year. In that case, all 4 proportions must be 1/4. So we could write the hypotheses in symbols as

H_0: $p_{\text{Jan-Mar}} = p_{\text{Apr-Jun}} = p_{\text{Jul-Sep}} = p_{\text{Oct-Dec}} = 1/4$
H_a: At least two of the proportions are not 1/4

STATE: We want to perform a test of

H_0: The birthdays of all NHL players are evenly distributed across the four quarters of the year.

H_a: The birthdays of all NHL players are not evenly distributed across the four quarters of the year.

No significance level was specified, so we'll use $\alpha = 0.05$.

PLAN: If the conditions are met, we will perform a chi-square test for goodness of fit.

- *Random:* The data came from a random sample of NHL players.
  - 10%: Because we are sampling without replacement, there must be at least $10(80) = 800$ NHL players. In the season when the data were collected, there were 879 NHL players.
- *Large Counts:* If birthdays are evenly distributed across the four quarters of the year, then the expected counts are all $80(1/4) = 20$. These counts are all at least 5.

DO:

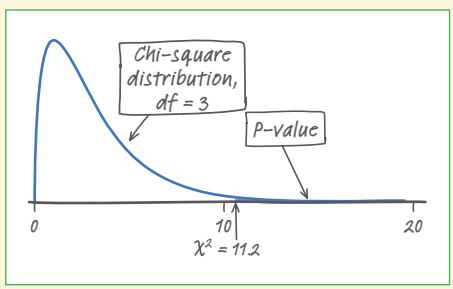


FIGURE 11.5 The *P*-value for the chi-square test for goodness of fit with $\chi^2 = 11.2$ and df = 3.

- *Test statistic:*

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

$$= \frac{(32 - 20)^2}{20} + \frac{(20 - 20)^2}{20} + \frac{(16 - 20)^2}{20} + \frac{(12 - 20)^2}{20}$$

$$= 7.2 + 0 + 0.8 + 3.2 = 11.2$$

- *P-value:* Figure 11.5 displays the *P*-value for this test as an area under the chi-square distribution with $4 - 1 = 3$ degrees of freedom. As the excerpt at right shows, $\chi^2 = 11.2$ corresponds to a *P*-value between 0.01 and 0.02.

Using Technology: Refer to the Technology Corner that follows the example. The calculator's χ^2 GOF-Test gives $\chi^2 = 11.2$ and *P*-value = 0.011 using df = 3.

CONCLUDE: Because the *P*-value, 0.011, is less than $\alpha = 0.05$, we reject H_0. We have convincing evidence that the birthdays of NHL players are not evenly distributed across the four quarters of the year.

	p		
df	**0.02**	**0.01**	**0.005**
2	7.82	9.21	10.60
3	9.84	11.34	12.84
4	11.67	13.28	14.86

For Practice *Try Exercise* 15

You can use your calculator to carry out the "Do" step for a chi-square test for goodness of fit. Remember that this comes with potential benefits and risks on the AP® exam.

26. TECHNOLOGY CORNER CHI-SQUARE TEST FOR GOODNESS OF FIT ON THE CALCULATOR

TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

You can use the TI-83/84 or TI-89 to perform the calculations for a chi-square test for goodness of fit. We'll use the data from the hockey and birthdays example to illustrate the steps.

1. Enter the counts.

- Enter the observed counts in L1/list1. Enter the expected counts in L2/list2.

Birthday	Observed	Expected
Jan–Mar	32	20
Apr–Jun	20	20
Jul–Sep	16	20
Oct–Dec	12	20

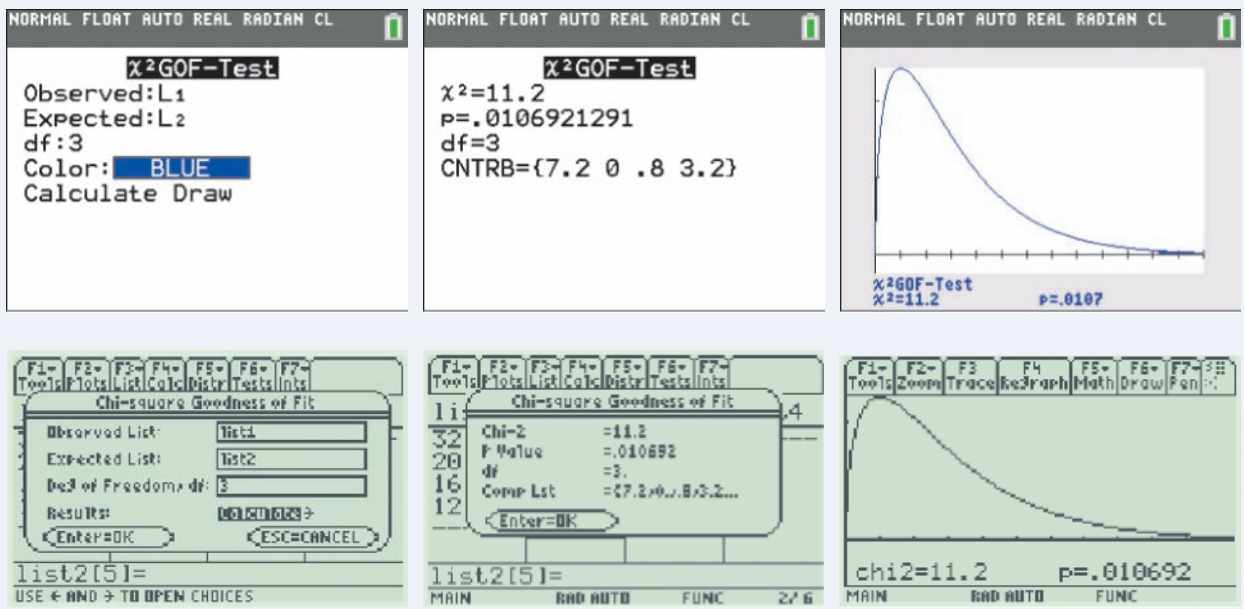
2. Perform a chi-square test for goodness of fit.

Note: Some older TI-83s and TI-84s don't have this test. TI-84 users can get this functionality by upgrading their operating systems.

TI-83/84: Press [STAT], arrow over to TESTS and choose χ^2GOF-Test... .

TI-89: In the Stats/List Editor APP, press [2nd] [F1] ([F6]) and choose Chi2GOF... .

Enter the inputs shown below. If you choose Calculate, you'll get a screen with the test statistic, *P*-value, and df. If you choose the Draw option, you'll get a picture of the appropriate chi-square distribution with the test statistic marked and shaded area corresponding to the *P*-value.



We'll discuss the CNTRB and Comp Lst results shortly.

AP® EXAM TIP You can use your calculator to carry out the mechanics of a significance test on the AP® exam. But there's a risk involved. If you just give the calculator answer with no work, and one or more of your values is incorrect, you will probably get no credit for the "Do" step. We recommend writing out the first few terms of the chi-square calculation followed by "...". This approach might help you earn partial credit if you enter a number incorrectly. Be sure to name the procedure (χ^2GOF-Test) and to report the test statistic ($\chi^2 = 11.2$), degrees of freedom (df = 3), and *P*-value (0.011).

Follow-up Analysis In the chi-square test for goodness of fit, we test the null hypothesis that a categorical variable has a specified distribution in the population of interest. If the sample data lead to a statistically significant result, we can conclude that our variable has a distribution different from the one stated. When this happens, start by examining which categories of the variable show large deviations between the observed and expected counts. Then look at the individual terms $\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$ that are added together to produce the test statistic χ^2. These **components** show which terms contribute most to the chi-square statistic.

Let's return to the hockey and birthdays example. The table of observed and expected counts for the 80 randomly selected NHL players is repeated below. We have added a column that shows the components of the chi-square test statistic. Looking at the counts, we see that there were many more players born in January through March than expected and far fewer players born in October through December than expected. The component for January to March birthdays made the largest contribution to the chi-square statistic. These results support Malcolm Gladwell's claim that NHL players are more likely to be born early in the year.

Birthday	Observed	Expected	O − E	$(O-E)^2/E$
Jan–Mar	32	20	12	7.2
Apr–Jun	20	20	0	0.0
Jul–Sep	16	20	−4	0.8
Oct–Dec	12	20	−8	3.2

Note: When we ran the chi-square test for goodness of fit on the calculator, a list of these individual components was stored. On the TI-83/84, the list is called CNTRB (for contribution). On the TI-89, it's called Comp Lst (component list).

CHECK YOUR UNDERSTANDING

Biologists wish to mate pairs of fruit flies having genetic makeup RrCc, indicating that each has one dominant gene (R) and one recessive gene (r) for eye color, along with one dominant (C) and one recessive (c) gene for wing type. Each offspring will receive one gene for each of the two traits from each parent. The following table, known as a *Punnett square*, shows the possible combinations of genes received by the offspring:

	Parent 2 passes on:			
Parent 1 passes on:	**RC**	**Rc**	**rC**	**rc**
RC	RRCC (x)	RRCc (x)	RrCC (x)	RrCc (x)
Rc	RRCc (x)	RRcc (y)	RrCc (x)	Rrcc (y)
rC	RrCC (x)	RrCc (x)	rrCC (z)	rrCc (z)
rc	RrCc (x)	Rrcc (y)	rrCc (z)	rrcc (w)

Any offspring receiving an R gene will have red eyes, and any offspring receiving a C gene will have straight wings. So based on this Punnett square, the biologists predict a ratio of 9 red-eyed, straight-winged (x):3 red-eyed, curly-winged (y):3 white-eyed, straight-winged (z):1 white-eyed, curly-winged (w) offspring.

To test their hypothesis about the distribution of offspring, the biologists mate a random sample of pairs of fruit flies. Of 200 offspring, 99 had red eyes and straight wings, 42 had red eyes and curly wings, 49 had white eyes and straight wings, and 10 had white eyes and curly wings. Do these data differ significantly from what the biologists have predicted? Carry out a test at the $\alpha = 0.01$ significance level.

Section 11.1 Summary

- A **one-way table** is often used to display the distribution of a single categorical variable for a sample of individuals.
- The **chi-square test for goodness of fit** tests the null hypothesis that a categorical variable has a specified distribution in the population of interest.
- This test compares the **observed count** in each category with the counts that would be expected if H_0 were true. The **expected count** for any category is found by multiplying the sample size by the proportion in each category according to the null hypothesis.
- The **chi-square statistic** is

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

  where the sum is over all possible categories.
- The conditions for performing a chi-square test for goodness of fit are:
  - **Random:** The data were produced by a well-designed random sample or randomized experiment.
    - **10%:** When sampling without replacement, check that the population is at least 10 times as large as the sample.
  - **Large Counts:** All expected counts must be at least 5.
- When the conditions are met, the sampling distribution of the statistic χ^2 can be modeled by a **chi-square distribution**.
- Large values of χ^2 are evidence against H_0 and in favor of H_a. The *P*-value is the area to the right of χ^2 under the chi-square distribution with degrees of freedom df = number of categories − 1.
- If the test finds a statistically significant result, consider doing a follow-up analysis that compares the observed and expected counts and that looks for the largest **components** of the chi-square statistic.

11.1 TECHNOLOGY CORNERS

TI-Nspire Instructions in Appendix B; HP Prime instructions on the book's Web site.

Section 11.1 Exercises

1. **Aw, nuts!** A company claims that each batch of its deluxe mixed nuts contains 52% cashews, 27% almonds, 13% macadamia nuts, and 8% brazil nuts. To test this claim, a quality-control inspector takes a random sample of 150 nuts from the latest batch. The one-way table below displays the sample data.

pg 681

Nut:	Cashew	Almond	Macadamia	Brazil
Count:	83	29	20	18

(a) State appropriate hypotheses for performing a test of the company's claim.

(b) Calculate the expected counts for each type of nut. Show your work.

2. **Roulette** Casinos are required to verify that their games operate as advertised. American roulette wheels have 38 slots—18 red, 18 black, and 2 green. In one casino, managers record data from a random sample of 200 spins of one of their American roulette wheels. The one-way table below displays the results.

Color:	Red	Black	Green
Count:	85	99	16

(a) State appropriate hypotheses for testing whether these data give convincing evidence that the distribution of outcomes on this wheel is not what it should be.

(b) Calculate the expected counts for each color. Show your work.

pg 683

3. **Aw, nuts!** Calculate the chi-square statistic for the data in Exercise 1. Show your work.

4. **Roulette** Calculate the chi-square statistic for the data in Exercise 2. Show your work.

5. **Aw, nuts!** Refer to Exercises 1 and 3.

(a) Confirm that the expected counts are large enough to use a chi-square distribution to calculate the *P*-value. What degrees of freedom should you use?

(b) Sketch a graph like Figure 11.4 (page 685) that shows the *P*-value.

(c) Use Table C to find the *P*-value. Then use your calculator's χ^2`cdf` command.

(d) What conclusion would you draw about the company's claimed distribution for its deluxe mixed nuts? Justify your answer.

6. **Roulette** Refer to Exercises 2 and 4.

(a) Confirm that the expected counts are large enough to use a chi-square distribution to calculate the *P*-value. What degrees of freedom should you use?

(b) Sketch a graph like Figure 11.4 (page 685) that shows the *P*-value.

(c) Use Table C to find the *P*-value. Then use your calculator's χ^2`cdf` command.

(d) What conclusion would you draw about whether or not the roulette wheel is operating correctly? Justify your answer.

7. **Birds in the trees** Researchers studied the behavior of birds that were searching for seeds and insects in an Oregon forest. In this forest, 54% of the trees were Douglas firs, 40% were ponderosa pines, and 6% were other types of trees. At a randomly selected time during the day, the researchers observed 156 red-breasted nuthatches: 70 were seen in Douglas firs, 79 in ponderosa pines, and 7 in other types of trees.[2] Do these data provide convincing evidence that nuthatches prefer particular types of trees when they're searching for seeds and insects?

8. **Seagulls by the seashore** Do seagulls show a preference for where they land? To answer this question, biologists conducted a study in an enclosed outdoor space with a piece of shore whose area was made up of 56% sand, 29% mud, and 15% rocks. The biologists chose 200 seagulls at random. Each seagull was released into the outdoor space on its own and observed until it landed somewhere on the piece of shore. In all, 128 seagulls landed on the sand, 61 landed in the mud, and 11 landed on the rocks. Do these data provide convincing evidence that seagulls show a preference for where they land?

9. **No chi-square** A school's principal wants to know if students spend about the same amount of time on homework each night of the week. She asks a random sample of 50 students to keep track of their homework time for a week. The following table displays the average amount of time (in minutes) students reported per night:

Night:	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Average time:	130	108	115	104	99	37	62

Explain carefully why it would not be appropriate to perform a chi-square test for goodness of fit using these data.

10. **No chi-square** The principal in Exercise 9 also asked the random sample of students to record whether they did all of the homework that was assigned on each of the five school days that week. Here are the data:

School day:	Monday	Tuesday	Wednesday	Thursday	Friday
No. who did homework:	34	29	32	28	19

Explain carefully why it would not be appropriate to perform a chi-square test for goodness of fit using these data.

11. **Benford's law** Faked numbers in tax returns, invoices, or expense account claims often display patterns that aren't present in legitimate records. Some patterns are obvious and easily avoided by a clever crook. Others are more subtle. It is a striking fact that the first digits of numbers in legitimate records often follow a model known as Benford's law.[3] Call the first digit of a randomly chosen record X for short. Benford's law gives this probability model for X (note that a first digit can't be 0):

First digit:	1	2	3	4	5	6	7	8	9
Probability:	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

A forensic accountant who is familiar with Benford's law inspects a random sample of 250 invoices from a company that is accused of committing fraud. The table below displays the sample data.

First digit:	1	2	3	4	5	6	7	8	9
Count:	61	50	43	34	25	16	7	8	6

(a) Are these data inconsistent with Benford's law? Carry out an appropriate test at the $\alpha = 0.05$ level to support your answer. If you find a significant result, perform a follow-up analysis.

(b) Describe a Type I error and a Type II error in this setting, and give a possible consequence of each. Which do you think is more serious?

12. **Housing** According to the Census Bureau, the distribution by ethnic background of the New York City population in a recent year was

Hispanic: 28% Black: 24% White: 35%
Asian: 12% Others: 1%

The manager of a large housing complex in the city wonders whether the distribution by race of the complex's residents is consistent with the population distribution. To find out, she records data from a random sample of 800 residents. The table below displays the sample data.[4]

Race:	Hispanic	Black	White	Asian	Other
Count:	212	202	270	94	22

Are these data significantly different from the city's distribution by race? Carry out an appropriate test at the $\alpha = 0.05$ level to support your answer. If you find a significant result, perform a follow-up analysis.

13. **Skittles** Statistics teacher Jason Molesky contacted Mars, Inc., to ask about the color distribution for Skittles candies. Here is an excerpt from the response he received: "The original flavor blend for the SKITTLES BITE SIZE CANDIES is lemon, lime, orange, strawberry and grape. They were chosen as a result of consumer preference tests we conducted. The flavor blend is 20 percent of each flavor."

(a) State appropriate hypotheses for a significance test of the company's claim.

(b) Find the expected counts for a bag of Skittles with 60 candies.

(c) How large a χ^2 statistic would you need to have significant evidence against the company's claim at the $\alpha = 0.05$ level? At the $\alpha = 0.01$ level?

(d) Create a set of observed counts for a bag with 60 candies that gives a *P*-value between 0.01 and 0.05. Show the calculation of your chi-square statistic.

14. **Is your random number generator working?** Use your calculator's RandInt function to generate 200 digits from 0 to 9 and store them in a list.

(a) State appropriate hypotheses for a chi-square test for goodness of fit to determine whether your calculator's random number generator gives each digit an equal chance to be generated.

(b) Carry out a test at the $\alpha = 0.05$ significance level.

For parts (c) and (d), assume that the students' random number generators are all working properly.

(c) What is the probability that a student who does this exercise will make a Type I error?

(d) Suppose that 25 students in an AP Statistics class independently do this exercise for homework. Find the probability that at least one of them makes a Type I error.

15. **What's your sign?** The University of Chicago's General Social Survey (GSS) is the nation's most important social science sample survey. For reasons known

pg 688

only to social scientists, the GSS regularly asks a random sample of people their astrological sign. Here are the counts of responses from a recent GSS:

Sign:	Aries	Taurus	Gemini	Cancer	Leo	Virgo
Count:	321	360	367	374	383	402
Sign:	**Libra**	**Scorpio**	**Sagittarius**	**Capricorn**	**Aquarius**	**Pisces**
Count:	392	329	331	354	376	355

If births are spread uniformly across the year, we expect all 12 signs to be equally likely. Do these data provide convincing evidence that all 12 signs are not equally likely? If you find a significant result, perform a follow-up analysis.

16. **Munching Froot Loops** Kellogg's Froot Loops cereal comes in six fruit flavors: orange, lemon, cherry, raspberry, blueberry, and lime. Charise poured out her morning bowl of cereal and methodically counted the number of cereal pieces of each flavor. Here are her data:

Flavor:	Orange	Lemon	Cherry	Raspberry	Blueberry	Lime
Count:	28	21	16	25	14	16

Do these data provide convincing evidence that Kellogg's Froot Loops do not contain an equal proportion of each flavor? If you find a significant result, perform a follow-up analysis.

17. **Mendel and the peas** Gregor Mendel (1822–1884), an Austrian monk, is considered the father of genetics. Mendel studied the inheritance of various traits in pea plants. One such trait is whether the pea is smooth or wrinkled. Mendel predicted a ratio of 3 smooth peas for every 1 wrinkled pea. In one experiment, he observed 423 smooth and 133 wrinkled peas. Assume that the conditions for inference were met. Carry out an appropriate test of the genetic model that Mendel predicted. What do you conclude?

18. **You say tomato** The paper "Linkage Studies of the Tomato" (*Transactions of the Canadian Institute*, 1931) reported the following data on phenotypes resulting from crossing tall cut-leaf tomatoes with dwarf potato-leaf tomatoes. We wish to investigate whether the following frequencies are consistent with genetic laws, which state that the phenotypes should occur in the ratio 9:3:3:1.

Phenotype:	Tall cut	Tall potato	Dwarf cut	Dwarf potato
Frequency:	926	288	293	104

Assume that the conditions for inference were met. Carry out an appropriate test of the proposed genetic model. What do you conclude?

Multiple choice: Select the best answer for Exercises 19 to 22.

Exercises 19 to 21 refer to the following setting. The manager of a high school cafeteria is planning to offer several new types of food for student lunches in the following school year. She wants to know if each type of food will be equally popular so she can start ordering supplies and making other plans. To find out, she selects a random sample of 100 students and asks them, "Which type of food do you prefer: Asian food, Mexican food, pizza, or hamburgers?" Here are her data:

Type of Food:	Asian	Mexican	Pizza	Hamburgers
Count:	18	22	39	21

19. An appropriate null hypothesis to test whether the food choices are equally popular is

(a) $H_0: \mu = 25$, where μ = the mean number of students that prefer each type of food.

(b) $H_0: p = 0.25$, where p = the proportion of all students who prefer Asian food.

(c) $H_0: n_A = n_M = n_P = n_H = 25$, where n_A is the number of students in the school who would choose Asian food, and so on.

(d) $H_0: p_A = p_M = p_P = p_H = 0.25$, where p_A is the proportion of students in the school who would choose Asian food, and so on.

(e) $H_0: \hat{p}_A = \hat{p}_M = \hat{p}_P = \hat{p}_H = 0.25$, where $\hat{p}_A$ is the proportion of students in the sample who chose Asian food, and so on.

20. The chi-square statistic is

(a) $\dfrac{(18-25)^2}{25} + \dfrac{(22-25)^2}{25} + \dfrac{(39-25)^2}{25} + \dfrac{(21-25)^2}{25}$

(b) $\dfrac{(25-18)^2}{18} + \dfrac{(25-22)^2}{22} + \dfrac{(25-39)^2}{39} + \dfrac{(25-21)^2}{21}$

(c) $\dfrac{(18-25)}{25} + \dfrac{(22-25)}{25} + \dfrac{(39-25)}{25} + \dfrac{(21-25)}{25}$

(d) $\dfrac{(18-25)^2}{100} + \dfrac{(22-25)^2}{100} + \dfrac{(39-25)^2}{100} + \dfrac{(21-25)^2}{100}$

(e) $\dfrac{(0.18-0.25)^2}{0.25} + \dfrac{(0.22-0.25)^2}{0.25} + \dfrac{(0.39-0.25)^2}{0.25} + \dfrac{(0.21-0.25)^2}{0.25}$

21. The P-value for a chi-square test for goodness of fit is 0.0129. Which of the following is the most appropriate conclusion?

(a) Because 0.0129 is less than $\alpha = 0.05$, reject H_0. There is convincing evidence that the food choices are equally popular.

(b) Because 0.0129 is less than $\alpha = 0.05$, reject H_0. There is not convincing evidence that the food choices are equally popular.

(c) Because 0.0129 is less than $\alpha = 0.05$, reject H_0. There is convincing evidence that the food choices are not equally popular.

(d) Because 0.0129 is less than $\alpha = 0.05$, fail to reject H_0. There is not convincing evidence that the food choices are equally popular.

(e) Because 0.0129 is less than $\alpha = 0.05$, fail to reject H_0. There is convincing evidence that the food choices are equally popular.

22. Which of the following is *false*?

(a) A chi-square distribution with k degrees of freedom is more right-skewed than a chi-square distribution with $k + 1$ degrees of freedom.

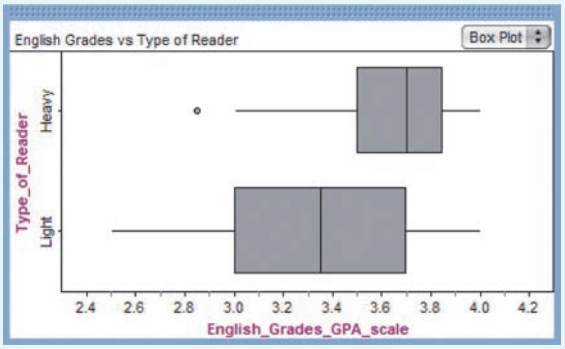
(b) A chi-square distribution never takes negative values.

(c) The degrees of freedom for a chi-square test is determined by the sample size.

(d) $P(\chi^2 > 10)$ is greater when df $= k + 1$ than when df $= k$.

(e) The area under a chi-square density curve is always equal to 1.

Exercises 23 through 25 refer to the following setting. Do students who read more books for pleasure tend to earn higher grades in English? The boxplots below show data from a simple random sample of 79 students at a large high school. Students were classified as light readers if they read fewer than 3 books for pleasure per year. Otherwise, they were classified as heavy readers. Each student's average English grade for the previous two marking periods was converted to a GPA scale where A+ = 4.3, A = 4.0, A− =3.7, B + =3.3, and so on.



23. **Reading and grades** (1.3) Write a few sentences comparing the distributions of English grades for light and heavy readers.

24. **Reading and grades** (10.2) Summary statistics for the two groups from Minitab are provided below.

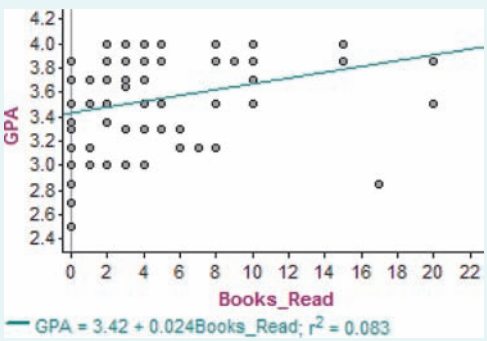
Type of reader	N	Mean	StDev	SE Mean
Heavy	47	3.640	0.324	0.047
Light	32	3.356	0.380	0.067

(a) Explain why it is acceptable to use two-sample t procedures in this setting.

(b) Construct and interpret a 95% confidence interval for the difference in the mean English grade for light and heavy readers.

(c) Does the interval in part (b) provide convincing evidence that reading more causes a difference in students' English grades? Justify your answer.

25. **Reading and grades** (3.2) The Fathom scatterplot below shows the number of books read and the English grade for all 79 students in the study. A least-squares regression line has been added to the graph.



(a) Interpret the meaning of the slope and y intercept in context.

(b) The student who reported reading 17 books for pleasure had an English GPA of 2.85. Find this student's residual and interpret this value in context.

(c) How strong is the relationship between English grades and number of books read? Give appropriate evidence to support your answer.

26. **Yahtzee** (5.3, 6.3) In the game of Yahtzee, 5 six-sided dice are rolled simultaneously. To get a Yahtzee, the player must get the same number on all 5 dice.

(a) Luis says that the probability of getting a Yahtzee in one roll of the dice is $\left(\frac{1}{6}\right)^5$. Explain why Luis is wrong.

(b) Nassir decides to keep rolling all 5 dice until he gets a Yahtzee. He is surprised when he still hasn't gotten a Yahtzee after 25 rolls. Should he be? Calculate an appropriate probability to support your answer.

11.2 Inference for Two-Way Tables

WHAT YOU WILL LEARN By the end of the section, you should be able to:

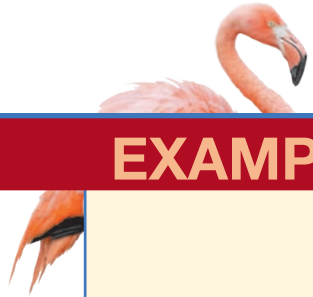
- Compare conditional distributions for data in a two-way table.
- State appropriate hypotheses and compute expected counts for a chi-square test based on data in a two-way table.
- Calculate the chi-square statistic, degrees of freedom, and *P*-value for a chi-square test based on data in a two-way table.
- Perform a chi-square test for homogeneity.
- Perform a chi-square test for independence.
- Choose the appropriate chi-square test.

The two-sample z procedures of Chapter 10 allow us to compare the proportions of successes in two populations or for two treatments. What if we want to compare more than two samples or groups? More generally, what if we want to compare the distributions of a single categorical variable across several populations or treatments? We need a new statistical test. The new test starts by presenting the data in a two-way table.

Two-way tables have more general uses than comparing distributions of a single categorical variable. As we saw in Section 1.1, they can be used to describe relationships between any two categorical variables. In this section, we will start by developing a test to determine whether the distribution of a categorical variable is the same for each of several populations or treatments. This test is called a **chi-square test for homogeneity**. Then we'll examine a related test to see whether there is convincing evidence of an association between the row and column variables in a two-way table. This test is known as a **chi-square test for independence**.

Comparing Distributions of a Categorical Variable

We'll start with an example involving a randomized experiment.



EXAMPLE

Does Background Music Influence What Customers Buy?

Comparing conditional distributions

Market researchers suspect that background music may affect the mood and buying behavior of customers. One study in a European restaurant compared three randomly assigned treatments: no music, French accordion music, and Italian string music. Under each condition, the researchers recorded the number of

customers who ordered French, Italian, and other entrees.[5] Here is a table that summarizes the data:

	Type of Music			
Entree ordered	None	French	Italian	Total
French	30	39	30	**99**
Italian	11	1	19	**31**
Other	43	35	35	**113**
Total	**84**	**75**	**84**	**243**

PROBLEM:

(a) Calculate the conditional distribution (in proportions) of entrees ordered for each treatment.

(b) Make an appropriate graph for comparing the conditional distributions in part (a).

(c) Write a few sentences comparing the distributions of entrees ordered under the three music treatments.

SOLUTION:

(a) When no music was playing, the distribution of entree orders was

$$\text{French: } \frac{30}{84} = 0.357 \quad \text{Italian: } \frac{11}{84} = 0.131 \quad \text{Other: } \frac{43}{84} = 0.512$$

When French accordion music was playing, the distribution of entree orders was

$$\text{French: } \frac{39}{75} = 0.520 \quad \text{Italian: } \frac{1}{75} = 0.013 \quad \text{Other: } \frac{35}{75} = 0.467$$

When Italian string music was playing, the distribution of entree orders was

$$\text{French: } \frac{30}{84} = 0.357 \quad \text{Italian: } \frac{19}{84} = 0.226 \quad \text{Other: } \frac{35}{84} = 0.417$$

(b) The bar graphs in Figure 11.6 compare the distributions of entrees ordered for each of the three music treatments.

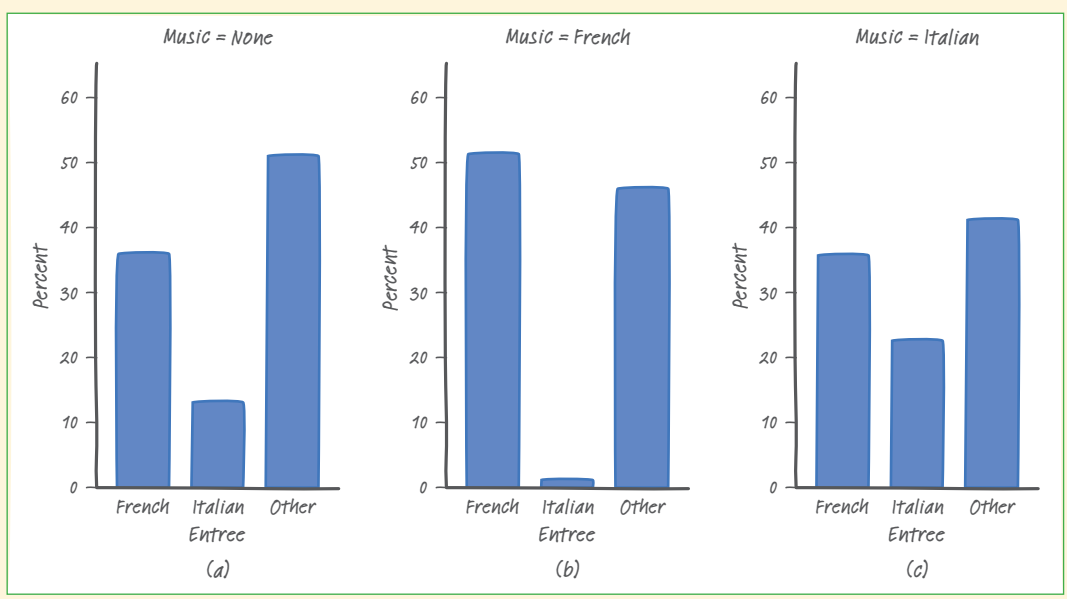


FIGURE 11.6 Bar graphs comparing the distributions of entrees ordered for different music conditions.

(c) The type of entree that customers order seems to differ considerably across the three music treatments. Orders of Italian entrees are very low (1.3%) when French music is playing but are higher when Italian music (22.6%) or no music (13.1%) is playing. French entrees seem popular in this restaurant, as they are ordered frequently under all music conditions but notably more often when French music is playing. For all three music treatments, the percent of Other entrees ordered was similar.

For Practice *Try Exercise* 27

The researchers in the restaurant study expected that music would influence customer orders, so the type of music played is the explanatory variable and the type of entree ordered is the response variable. A good general strategy is to compare the conditional distributions of the response variable for each value of the explanatory variable. That's why we compared the conditional distributions of entrees ordered for each type of music played.

It is common practice to describe a two-way table by its number of rows and columns (not including totals). For instance, the data in the previous example were given in a 3×3 table. The following Check Your Understanding involves a 3×2 table.

CHECK YOUR UNDERSTANDING

The Pennsylvania State University has its main campus in the town of State College and more than 20 smaller "commonwealth campuses" around the state. The Penn State Division of Student Affairs polled separate random samples of undergraduates from the main campus and commonwealth campuses about their use of online social networking. Facebook was the most popular site, with more than 80% of students having an account. Here is a comparison of Facebook use by undergraduates at the main campus and commonwealth campuses who have a Facebook account:[6]

Use Facebook	Main campus	Commonwealth
Several times a month or less	55	76
At least once a week	215	157
At least once a day	640	394
Total Facebook users	**910**	**627**

1. Calculate the conditional distribution (in proportions) of Facebook use for each campus setting.
2. Why is it important to compare proportions rather than counts in Question 1?
3. Make a bar graph that compares the two conditional distributions. What are the most important differences in Facebook use between the two campus settings?

Stating Hypotheses

The null hypothesis in the restaurant example is

H_0: There is *no difference* in the true distributions of entrees ordered at this restaurant when no music, French accordion music, or Italian string music is played.

If the null hypothesis is true, the observed differences in the distributions of entrees ordered for the three groups are due to the chance involved in the random assignment of treatments. The alternative hypothesis says that there *is* a difference but does not specify the nature of that difference:

H_a: There is a difference in the true distributions of entrees ordered at this restaurant when no music, French accordion music, or Italian string music is played.

Any difference among the three true distributions of entrees ordered when no music, French accordion music, or Italian string music is played means that the null hypothesis is false and the alternative hypothesis is true. The alternative hypothesis is not one-sided or two-sided. We might call it "many-sided" because it allows any kind of difference.

With only the methods we already know, we might start by comparing the proportions of French entrees ordered when no music and French accordion music are played. We could similarly compare other pairs of proportions, ending up with many tests and many *P*-values. This is a bad idea. The *P*-values belong to each test separately, not to the collection of all the tests together.

	Type of Music			
Entree ordered	**None**	**French**	**Italian**	**Total**
French	30	39	30	**99**
Italian	11	1	19	**31**
Other	43	35	35	**113**
Total	**84**	**75**	**84**	**243**

When we do many individual tests or construct many confidence intervals, the individual P-values and confidence levels don't tell us how confident we can be in all the inferences taken together. Because of this, it's cheating to pick out one large difference from the two-way table and then perform a significance test as if it were the only comparison we had in mind. For example, the proportions of French entrees ordered under the no music and French accordion music treatments are 30/84 = 0.357 and 39/75 = 0.520, respectively. A two-sample z test shows that the difference between the proportions is statistically significant ($z = 2.06$, $P = 0.039$) if we make just this one comparison.

But we could also pick a comparison that is not significant. For example, the proportions of Italian entrees ordered for the no music and Italian string music treatments are 11/84 = 0.131 and 19/84 = 0.226, respectively. These two proportions do not differ significantly ($z = 1.61$, $P = 0.107$). Individual comparisons can't tell us whether the three *distributions* of the categorical variable (in this case, type of entree ordered) are significantly different.

The problem of how to do many comparisons at once with an overall measure of confidence in all our conclusions is common in statistics. This is the problem of **multiple comparisons.** Statistical methods for dealing with multiple comparisons usually have two parts:

1. An *overall test* to see if there is convincing evidence of any differences among the parameters that we want to compare.
2. A detailed *follow-up analysis* to decide which of the parameters differ and to estimate how large the differences are.

The overall test uses the familiar chi-square statistic. But in this new setting the test will be used to compare the distribution of a categorical variable for several populations or treatments. The follow-up analysis can be quite elaborate. We will concentrate on the overall test and do a follow-up analysis only when the observed results are statistically significant.

Expected Counts and the Chi-Square Statistic

A chi-square test for homogeneity begins with the hypotheses

H_0: There is no difference in the distribution of a categorical variable for several populations or treatments.

H_a: There is a difference in the distribution of a categorical variable for several populations or treatments.

It would also be correct to state the null hypothesis as H_0: The distribution of a categorical variable is the same for each of several populations or treatments. We prefer the "no difference" wording because it's more consistent with the language we used in the significance tests of Chapter 10.

To perform a test, we compare the observed counts in a two-way table with the counts we would expect if H_0 were true. Finding the expected counts is not that difficult, as the following example illustrates.

EXAMPLE Does Background Music Influence What Customers Buy?

Computing expected counts

The null hypothesis in the restaurant experiment is that there's no difference in the distribution of entrees ordered when no music, French accordion music, or Italian string music is played. To find the expected counts, we start by assuming that H_0 is true. We can see from the two-way table that 99 of the 243 entrees ordered during the study were French.

Observed Counts

	Type of Music			
Entree ordered	**None**	**French**	**Italian**	**Total**
French	30	39	30	**99**
Italian	11	1	19	**31**
Other	43	35	35	**113**
Total	**84**	**75**	**84**	**243**

If the specific type of music that's playing has no effect on entree orders, the proportion of French entrees ordered under each music condition should be 99/243 = 0.4074. For instance, there were 84 total entrees ordered when no music was playing. We would expect

$$84 \cdot \frac{99}{243} = 84(0.4074) = 34.22$$

Although any count of entrees ordered must be a whole number, an expected count need not be. The expected count gives the average number of entrees ordered if H_0 is true and the random assignment process is repeated many times.

of those entrees to be French, on average. The expected counts of French entrees ordered under the other two music conditions can be found in a similar way:

French music: $75(0.4074) = 30.56$ Italian music: $84(0.4074) = 34.22$

We repeat the process to find the expected counts for the other two types of entrees. The overall proportion of Italian entrees ordered during the study was $31/243 = 0.1276$. So the expected counts of Italian entrees ordered under each treatment are

No music: $84(0.1276) = 10.72$ French music: $75(0.1276) = 9.57$

Italian music: $84(0.1276) = 10.72$

The overall proportion of Other entrees ordered during the experiment was $113/243 = 0.465$. So the expected counts of Other entrees ordered for each treatment are

No music: $84(0.465) = 39.06$ French music: $75(0.465) = 34.88$

Italian music: $84(0.465) = 39.06$

The following table summarizes the expected counts for all three treatments. Note that the values for no music and Italian music are the same because 84 total entrees were ordered under each condition. We can check our work by adding the expected counts to obtain the row and column totals, as in the table. These should be the same as those in the table of observed counts except for small round-off errors, such as 75.01 rather than 75 for the total number of French entrees ordered.

Expected Counts

	Type of Music			
Entree ordered	**None**	**French**	**Italian**	**Total**
French	34.22	30.56	34.22	**99**
Italian	10.72	9.57	10.72	**31**
Other	39.06	34.88	39.06	**113**
Total	**84**	**75**	**84**	**243**

Let's take a look at the two-way table from the restaurant study one more time. In the example, we found the expected count of French entrees ordered when no music was playing as follows:

$$84 \cdot \frac{99}{243} = 34.22$$

Observed Counts

	Type of Music			
Entree ordered	**None**	**French**	**Italian**	**Total**
French	30	39	30	**99**
Italian	11	1	19	**31**
Other	43	35	35	**113**
Total	**84**	**75**	**84**	**243**

We marked the three numbers used in this calculation in the table. These values are the row total for French entrees ordered, the column total for entrees ordered

when no music was playing, and the table total of entrees ordered during the experiment. We can rewrite the original calculation as

$$\frac{84 \cdot 99}{243} = \frac{99 \cdot 84}{243} = 34.22$$

This suggests a more general formula for the expected count in any cell of a two-way table:

$$\frac{\text{row total} \cdot \text{column total}}{\text{table total}}$$

FINDING EXPECTED COUNTS

When H_0 is true, the expected count in any cell of a two-way table is

$$\text{expected count} = \frac{\text{row total} \cdot \text{column total}}{\text{table total}}$$

All the expected counts in the restaurant study are at least 5. This satisfies the Large Counts condition. The Random condition is met because the treatments were assigned at random. We don't need to check the 10% condition because the researchers were not sampling without replacement from some population of interest. They just performed an experiment using customers who happened to be in the restaurant at the time.

CONDITIONS FOR PERFORMING A CHI-SQUARE TEST FOR HOMOGENEITY

- **Random:** The data come from independent random samples or from the groups in a randomized experiment.
  - **10%:** When sampling without replacement, check that $n \leq \frac{1}{10}N$.
- **Large Counts:** All *expected* counts are at least 5.

Just as we did with the chi-square test for goodness of fit, we compare the observed counts with the expected counts using the statistic

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

This time, the sum is over all cells (not including the totals!) in the two-way table.

EXAMPLE Does Background Music Influence What Customers Buy?

The chi-square statistic

PROBLEM: The tables below show the observed and expected counts for the restaurant experiment. Calculate the chi-square statistic. Show your work.

Observed Counts

	Type of Music			
Entree ordered	None	French	Italian	Total
French	30	39	30	**99**
Italian	11	1	19	**31**
Other	43	35	35	**113**
Total	**84**	**75**	**84**	**243**

Expected Counts

	Type of Music			
Entree ordered	None	French	Italian	Total
French	34.22	30.56	34.22	**99**
Italian	10.72	9.57	10.72	**31**
Other	39.06	34.88	39.06	**113**
Total	**84**	**75**	**84**	**243**

AP® EXAM TIP In the "Do" step, you aren't required to show every term in the chi-square statistic. Writing the first few terms of the sum followed by ". . ." is considered as "showing work." We suggest that you do this and then let your calculator tackle the computations.

SOLUTION: For French entrees with no music, the observed count is 30 orders and the expected count is 34.22. The contribution to the χ^2 statistic for this cell is

$$\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} = \frac{(30 - 34.22)^2}{34.22} = 0.52$$

The χ^2 statistic is the sum of nine such terms:

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} = \frac{(30 - 34.22)^2}{34.22} + \frac{(39 - 30.56)^2}{30.56} + \cdots + \frac{(35 - 39.06)^2}{39.06}$$

$$= 0.52 + 2.33 + \cdots + 0.42 = 18.28$$

For Practice *Try Exercise* 29

As in the test for goodness of fit, you should think of the chi-square statistic χ^2 as a measure of how much the observed counts deviate from the expected counts. Once again, large values of χ^2 are evidence against H_0 and in favor of H_a. The P-value measures the strength of this evidence. When conditions are met, P-values for a chi-square test for homogeneity come from a chi-square distribution with df = (number of rows − 1) × (number of columns − 1).

EXAMPLE Does Background Music Influence What Customers Buy?

P-value and conclusion

Earlier, we started a significance test of

H_0: There is no difference in the true distributions of entrees ordered at this restaurant when no music, French accordion music, or Italian string music is played.

H_a: There is a difference in the true distributions of entrees ordered at this restaurant when no music, French accordion music, or Italian string music is played.

Observed Counts				
	Type of Music			
Entree ordered	None	French	Italian	Total
French	30	39	30	**99**
Italian	11	1	19	**31**
Other	43	35	35	**113**
Total	**84**	**75**	**84**	**243**

We already checked that the conditions are met. Our calculated test statistic is $\chi^2 = 18.28$.

PROBLEM:

(a) Use Table C to find the *P*-value. Then use your calculator's χ^2cdf command.

(b) Interpret the *P*-value from the calculator in context.

(c) What conclusion would you draw? Justify your answer.

SOLUTION:

(a) Because the two-way table that summarizes the data from the study has three rows and three columns, we use a chi-square distribution with df $= (3 - 1)(3 - 1) = 4$ to find the *P*-value.

- *Table:* Look at the df $= 4$ row in Table C. The calculated value $\chi^2 = 18.28$ lies between the critical values 16.42 and 18.47. The corresponding *P*-value is between 0.001 and 0.0025.

	P	
df	.0025	.001
4	16.42	18.47

- *Calculator:* The command χ^2cdf(18.28,10000,4) gives 0.0011.

(b) Assuming that there is no difference in the true distributions of entrees ordered in this restaurant when no music, French accordion music, or Italian string music is played, there is a 0.0011 probability of observing a difference in the distributions of entrees ordered among the three treatment groups as large as or larger than the one in this study.

(c) Because the *P*-value, 0.0011, is less than our default $\alpha = 0.05$ significance level, we reject H_0. We have convincing evidence of a difference in the true distributions of entrees ordered at this restaurant when no music, French accordion music, or Italian string music is played. Furthermore, the random assignment allows us to say that the difference is caused by the music that's played.

For Practice *Try Exercise* 31

CHECK YOUR UNDERSTANDING

In the previous Check Your Understanding (page 699), we presented data on the use of Facebook by two randomly selected groups of Penn State students. Here are the data once again.

Use Facebook	Main campus	Commonwealth
Several times a month or less	55	76
At least once a week	215	157
At least once a day	640	394
Total Facebook users	**910**	**627**

Do these data provide convincing evidence of a difference in the distributions of Facebook use among students in the two campus settings?

1. State appropriate null and alternative hypotheses for a significance test to help answer this question.
2. Calculate the expected counts. Show your work.
3. Calculate the chi-square statistic. Show your work.
4. Use Table C to find the *P*-value. Then use your calculator's χ^2cdf command.
5. Interpret the *P*-value from the calculator in context.
6. What conclusion would you draw? Justify your answer.

Calculating the expected counts and then the chi-square statistic by hand is a bit time-consuming. As usual, technology saves time and gets the arithmetic right.

27. TECHNOLOGY CORNER

CHI-SQUARE TESTS FOR TWO-WAY TABLES ON THE CALCULATOR

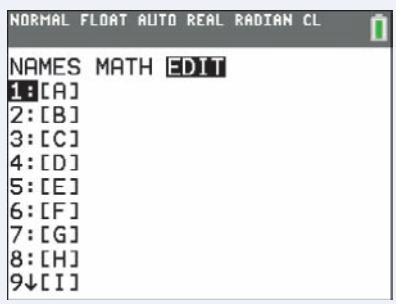
TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

You can use the TI-83/84 or TI-89 to perform calculations for a chi-square test for homogeneity. We'll use the data from the restaurant study to illustrate the process.

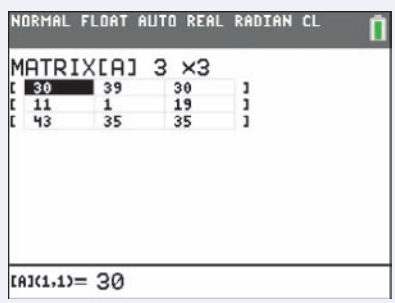
1. Enter the observed counts in matrix [A].

TI-83/84

- Press 2nd x^{-1} (MATRIX), arrow to EDIT, and choose A.
- Enter the dimensions of the matrix: 3 × 3.

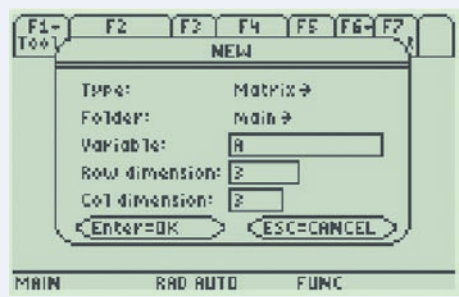


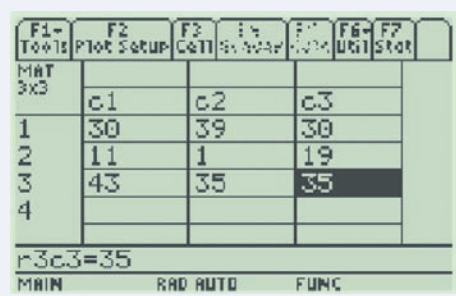
- Enter the observed counts from the two-way table in the same locations in the matrix.



TI-89

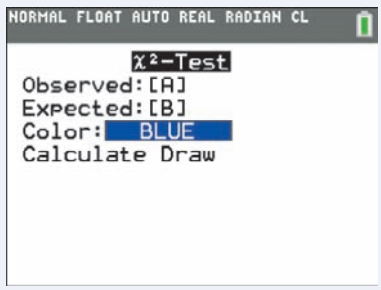
- Press APPS, select Data/Matrix Editor and then New. . . .
- Adjust your settings to match those shown.



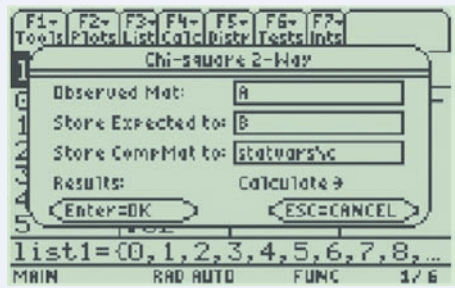


2. Specify the chi-square test, the matrix where the observed counts are found, and the matrix where the expected counts will be stored.

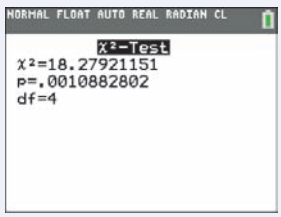
- Press STAT, arrow to TESTS, and choose χ^2-Test.
- Adjust your settings as shown.

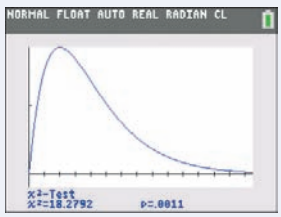


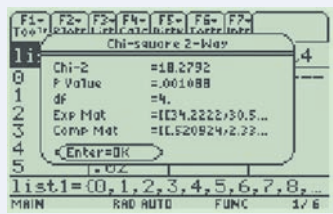
- In the Statistics/List Editor, press 2nd F1 ([F6]), and choose Chi2 2-way....
- Adjust your settings as shown.

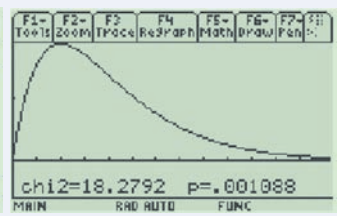


3. Choose "Calculate" or "Draw" to carry out the test. If you choose "Calculate," you should get the test statistic, *P*-value, and df shown below. If you specify "Draw," the chi-square distribution with 4 degrees of freedom will be drawn, the area in the tail will be shaded, and the *P*-value will be displayed.



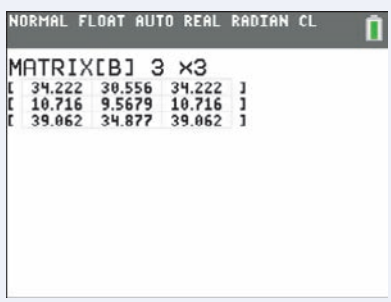




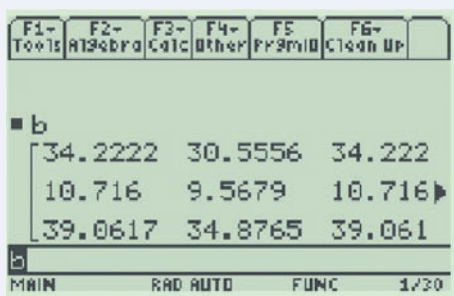


4. To see the expected counts, go to the home screen and ask for a display of the matrix [B].

- Press 2nd x^{-1} (MATRIX), arrow to EDIT, and choose [B].



- Press 2nd (−) (Var-LINK) and choose B.



AP® EXAM TIP You can use your calculator to carry out the mechanics of a significance test on the AP® exam. But there's a risk involved. If you just give the calculator answer with no work, and one or more of your values is incorrect, you will probably get no credit for the "Do" step. We recommend writing out the first few terms of the chi-square calculation followed by ". . . ". This approach might help you earn partial credit if you enter a number incorrectly. Be sure to name the procedure (χ^2-Test for homogeneity) and to report the test statistic ($\chi^2 = 18.279$), degrees of freedom (df = 4), and *P*-value (0.0011).

The Chi-Square Test for Homogeneity

In Section 11.1, we used a chi-square test for goodness of fit to test a hypothesized model for the distribution of a categorical variable. Our *P*-values came from a chi-square distribution with df = the number of categories − 1. When the Random, 10%, and Large Counts conditions are met, the χ^2 statistic calculated from

a two-way table can be used to perform a test of H_0: There is no difference in the distribution of a categorical variable for several populations or treatments. This new procedure is known as a chi-square test for homogeneity.

This test is also known as a chi-square test for homogeneity of proportions. We prefer the simpler name.

CHI-SQUARE TEST FOR HOMOGENEITY

Suppose the conditions are met. You can use the **chi-square test for homogeneity** to test

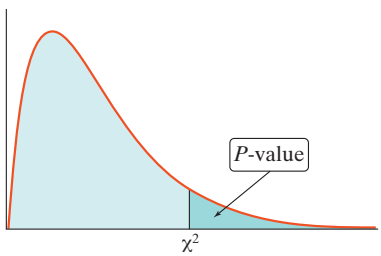
H_0: There is no difference in the distribution of a categorical variable for several populations or treatments.

H_a: There is a difference in the distribution of a categorical variable for several populations or treatments.

Start by finding the expected counts. Then calculate the chi-square statistic

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

where the sum is over all cells (not including totals) in the two-way table. If H_0 is true, the χ^2 statistic has approximately a chi-square distribution with degrees of freedom = (number of rows − 1)(number of columns − 1). The P-value is the area to the right of χ^2 under the corresponding chi-square density curve.



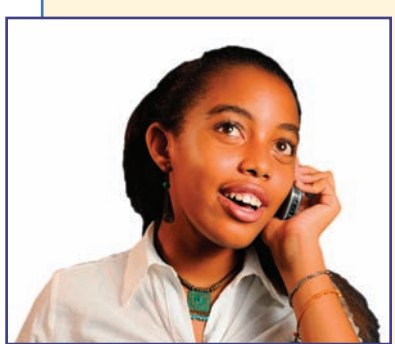
Let's look at an example of a chi-square test for homogeneity from start to finish. As usual, we follow the four-step process when performing a significance test.

EXAMPLE

Are Cell-Only Telephone Users Different?

STEP 4

The chi-square test for homogeneity



Random digit dialing telephone surveys used to exclude cell phone numbers. If the opinions of people who have only cell phones differ from those of people who have landline service, the poll results may not represent the entire adult population. The Pew Research Center interviewed separate random samples of cell-only and landline telephone users who were less than 30 years old. Here's what the Pew survey found about how these people describe their political party affiliation:[7]

	Cell-only sample	Landline sample
Democrat or lean Democratic	49	47
Refuse to lean either way	15	27
Republican or lean Republican	32	30
Total	**96**	**104**

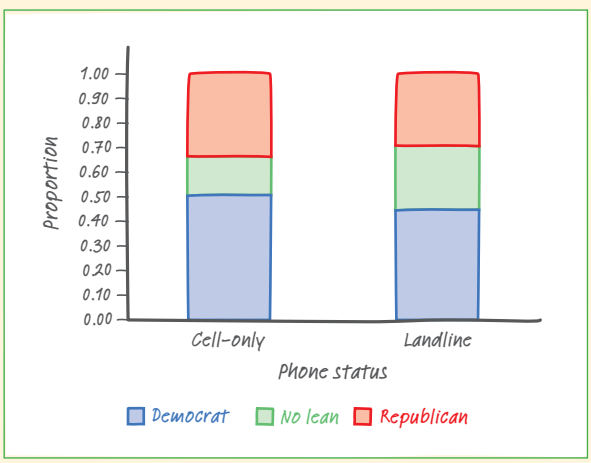
PROBLEM:

(a) Compare the distributions of political party affiliation for cell-only and landline phone users.

(b) Do these data provide convincing evidence at the $\alpha = 0.05$ level that the distribution of party affiliation differs in the under-30 cell-only and landline user populations?

SOLUTION:

(a) Because the sample sizes are different, we should compare the proportions of individuals in each political affiliation category in the two samples. The table below shows the conditional distributions of political party affiliation for cell-only and landline phone users. We made a segmented bar graph to compare these two distributions. Cell-only users appear slightly more likely to declare themselves as Democrats or Republicans than people who have landlines. People with landlines seem much more likely to say they don't lean Democratic or Republican than those who use only cell phones.



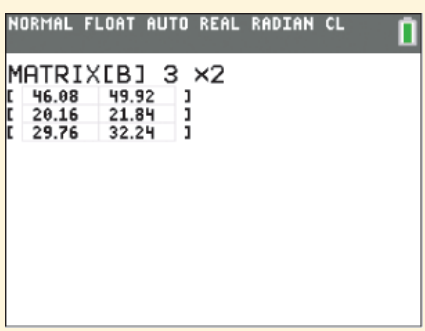
	Phone Status	
Political affiliation	**Cell only**	**Landline**
Democrat	0.51	0.45
No lean	0.16	0.26
Republican	0.33	0.29

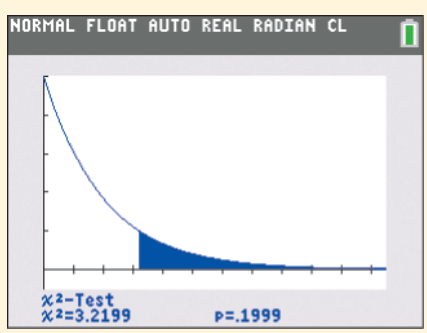
(b) **STATE:** We want to perform a test of

H_0: There is no difference in the distribution of party affiliation in the under-30 cell-only and landline populations.

H_a: There is a difference in the distribution of party affiliation in the under-30 cell-only and landline populations.

at the $\alpha = 0.05$ level.





PLAN: If conditions are met, we should use a chi-square test for homogeneity.

- *Random:* The data came from independent random samples of 96 cell-only and 104 landline users.
  - 10%: Sampling without replacement was used, so there need to be at least $10(96) = 960$ cell-only users under age 30 and at least $10(104) = 1040$ landline users under age 30. This is safe to assume.
- *Large Counts:* We followed the steps in the Technology Corner on page 706 to get the expected counts. The calculator screen shot confirms that all expected counts are at least 5.

DO: A chi-square test on the calculator gave

- *Test statistic:*

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

$$= \frac{(49 - 46.08)^2}{46.08} + \frac{(47 - 49.92)^2}{49.92} + \cdots = 3.22$$

- *P-value:* Using df = (number of rows − 1)(number of columns − 1) = (3 − 1)(2 − 1) = 2, the *P*-value is 0.1999.

CONCLUDE: Because our *P*-value, 0.1999, is greater than $\alpha = 0.05$, we fail to reject H_0. There is not convincing evidence that the distribution of party affiliation differs in the under-30 cell-only and landline user populations.

For Practice *Try Exercise* 33

Follow-up Analysis The chi-square test for homogeneity allows us to compare the distribution of a categorical variable for any number of populations or treatments. If the test allows us to reject the null hypothesis of no difference, we may want to do a follow-up analysis that examines the differences in detail. Start by examining which cells in the two-way table show large deviations between the observed and expected counts. Then look at the individual components $\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$ to see which terms contribute most to the chi-square statistic.

Our earlier restaurant study found significant differences among the true distributions of entrees ordered under each of the three music conditions. We entered the two-way table for the study into Minitab software and requested a chi-square test. The output appears in Figure 11.7. Minitab repeats the two-way table of observed counts and puts the expected count for each cell below the observed count. Finally, the software prints the 9 individual components that contribute to the χ^2 statistic.

Chi-Square Test: None, French, Italian

Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

	None	French	Italian	Total
1	30	39	30	99
	34.22	30.56	34.22	
	0.521	2.334	0.521	
2	11	1	19	31
	10.72	9.57	10.72	
	0.008	7.672	6.404	
3	43	35	35	113
	39.06	34.88	39.06	
	0.397	0.000	0.422	
Total	84	75	84	243

Chi-Sq = 18.279, DF = 4, P-Value = 0.001

FIGURE 11.7 Minitab output for the two-way table in the restaurant study. The output gives the observed counts, the expected counts, and the individual components of the chi-square statistic.

Looking at the output, we see that just two of the nine components that make up the chi-square statistic contribute about 14 (almost 77%) of the total $\chi^2 = 18.28$. Comparing the observed and expected counts in these two cells, we see that orders of Italian entrees are much below expectation when French music is playing and well above expectation when Italian music is playing. We are led to a specific conclusion: orders of Italian entrees are strongly affected by Italian and French music. More advanced methods provide tests and confidence intervals that make this follow-up analysis more complete.

THINK ABOUT IT

What if we want to compare several proportions? Many studies involve comparing the proportion of successes for each of several populations or treatments. The two-sample z test from Chapter 10 allows us to test the null hypothesis $H_0: p_1 = p_2$, where p_1 and p_2 are the true proportions of successes for the two populations or treatments. The chi-square test for homogeneity allows us to test $H_0: p_1 = p_2 = \cdots = p_k$. This null hypothesis says that there is no difference in the proportions of successes for the k populations or treatments. The alternative hypothesis is H_a: at least two of the p_i's are different. Many students *incorrectly state* H_a as "all the proportions are different." Think about it this way: the opposite of "all the proportions are equal" is "some of the proportions are not equal."

CHECK YOUR UNDERSTANDING

Canada has universal health care. The United States does not but often offers more elaborate treatment to patients with access. How do the two systems compare in treating heart attacks? Researchers compared random samples of U.S. and Canadian heart attack patients. One key outcome was the patients' own assessment of their quality of life relative to what it had been before the heart attack. Here are the data for the patients who survived a year:

Quality of life	Canada	United States
Much better	75	541
Somewhat better	71	498
About the same	96	779
Somewhat worse	50	282
Much worse	19	65
Total	**311**	**2165**

1. Construct an appropriate graph to compare the distributions of opinion about quality of life among heart attack patients in Canada and the United States.

2. Is there a significant difference between the two distributions of quality-of-life ratings? Carry out an appropriate test at the $\alpha = 0.01$ level.

Relationships between Two Categorical Variables

Two-way tables can arise in several ways. The restaurant experiment compared entrees ordered for three music treatments. The phone use and political party affiliation observational study compared independent random samples from the cell-only and landline user populations. In both cases, we are comparing the distributions of a categorical variable for several populations or treatments. We use the chi-square test for homogeneity to perform inference in such settings.

Another common situation that leads to a two-way table is when a *single* random sample of individuals is chosen from a *single* population and then classified based on two categorical variables. In that case, our goal is to analyze the relationship between the variables. The next example describes a study of this type.

EXAMPLE

Angry People and Heart Disease

Relationships between two categorical variables

A study followed a random sample of 8474 people with normal blood pressure for about four years.[8] All the individuals were free of heart disease at the beginning of the study. Each person took the Spielberger Trait Anger Scale test, which measures how prone a person is to sudden anger. Researchers also recorded whether each individual developed coronary heart disease (CHD). This includes people who had

heart attacks and those who needed medical treatment for heart disease. Here is a two-way table that summarizes the data:

	Low anger	Moderate anger	High anger	Total
CHD	53	110	27	**190**
No CHD	3057	4621	606	**8284**
Total	**3110**	**4731**	**633**	**8474**

PROBLEM:

(a) Is this an observational study or an experiment? Justify your answer.

(b) Make a well-labeled bar graph that compares CHD rates for the different anger levels. Describe what you see.

SOLUTION:

(a) This is an observational study. Researchers did not deliberately impose any treatments. They just recorded data about two variables—anger level and whether or not the person developed CHD—for each randomly chosen individual.

(b) In this setting, anger level is the explanatory variable and whether or not a person gets heart disease is the response variable. So we compare the percents of people who did and did not get heart disease in each of the three anger categories:

	CHD	no CHD
Low anger:	$\frac{53}{3110} = 0.0170 = 1.70\%$	$\frac{3057}{3110} = 0.9830 = 98.30\%$
Moderate anger:	$\frac{110}{4731} = 0.0233 = 2.33\%$	$\frac{4621}{4731} = 0.9767 = 97.67\%$
High anger:	$\frac{27}{633} = 0.0427 = 4.27\%$	$\frac{606}{633} = 0.9573 = 95.73\%$

The bar graph in Figure 11.8 shows the percent of people in each of the three anger categories who developed CHD. There is a clear trend: as the anger score increases, so does the percent who suffer heart disease. A much higher percent of people in the high anger category developed CHD (4.27%) than in the moderate (2.33%) and low (1.70%) anger categories.

FIGURE 11.8 Bar graph comparing the percents of people in each anger category who got coronary heart disease (CHD).

For Practice *Try Exercise* **41**

Anger rating on the Spielberger scale is a categorical variable that takes three possible values: low, medium, and high. Whether or not someone gets heart disease is also a categorical variable. The two-way table in the example shows the relationship between anger rating and heart disease for a random sample of

8474 people. Do these data provide convincing evidence of an *association* between the variables in the larger population? To answer that question, we work with a new significance test.

The Chi-Square Test for Independence

We often gather data from a random sample and arrange them in a two-way table to see if two categorical variables are associated. The sample data are easy to investigate: turn them into percents and look for a relationship between the variables. Is the association in the sample evidence of an association between these variables in the entire population? Or could the sample association easily arise just from the luck of random sampling? This is a question for a significance test.

Our null hypothesis is that there is *no association* between the two categorical variables in the population of interest. The alternative hypothesis is that there *is* an association between the variables. For the observational study of anger level and coronary heart disease, we want to test the hypotheses

H_0: There is no association between anger level and heart-disease status in the population of people with normal blood pressure.

H_a: There is an association between anger level and heart-disease status in the population of people with normal blood pressure.

No association between two variables means that knowing the value of one variable does not help us predict the value of the other. That is, the variables are *independent*. An equivalent way to state the hypotheses is therefore

We could substitute the word "dependent" in place of "not independent" in the alternative hypothesis. We'll avoid this practice, however, because saying that two variables are dependent sounds too much like saying that changes in one variable *cause* changes in the other.

H_0: Anger and heart-disease status are independent in the population of people with normal blood pressure.

H_a: Anger and heart-disease status are not independent in the population of people with normal blood pressure.

As with the two previous types of chi-square tests, we begin by comparing the observed counts in a two-way table with the expected counts if H_0 is true.

EXAMPLE

Angry People and Heart Disease

Finding expected counts

The null hypothesis is that there is no association between anger level and heart-disease status in the population of interest. If we assume that H_0 is true, then anger level and CHD status are independent. We can find the expected cell counts in the two-way table using the definition of independent events from Chapter 5: $P(A \mid B) = P(A)$. The chance process here is randomly selecting a person and recording his or her anger level and CHD status.

	Low anger	Moderate anger	High anger	Total
CHD	53	110	27	**190**
No CHD	3057	4621	606	**8284**
Total	**3110**	**4731**	**633**	**8474**

Let's start by considering the events "CHD" and "low anger." We see from the two-way table that 190 of the 8474 people in the study had CHD. If we imagine choosing one of these people at random, $P(\text{CHD}) = 190/8474$. Because anger level and CHD status are independent, knowing that the selected individual is low anger does not change the probability that this person develops CHD. That is to say, $P(\text{CHD} \mid \text{low anger}) = P(\text{CHD}) = 190/8474 = 0.02242$.

Of the 3110 low-anger people in the study, we'd expect

$$3110 \cdot \frac{190}{8474} = 3110(0.02242) = 69.73$$

to get CHD. You can see that the general formula we developed earlier for a test for homogeneity applies in this situation also:

$$\text{expected count} = \frac{\text{row total} \cdot \text{column total}}{\text{table total}} = \frac{190 \cdot 3110}{8474} = 69.73$$

To find the expected count in the "low anger, no CHD" cell, we begin by noting that $P(\text{no CHD}) = 8284/8474 = 0.97758$ for a randomly selected person in the study. Of the 3110 low-anger people in the study, we would expect

$$3110 \cdot \frac{8284}{8474} = 3110(0.97758) = 3040.27$$

to not develop CHD.

We find the expected counts for the remaining cells in the two-way table in a similar way.

CHD, Low	CHD, Moderate	CHD, High
$3110(0.02242) = 69.73$	$4731(0.02242) = 106.08$	$633(0.02242) = 14.19$
no CHD, Low	**no CHD, Moderate**	**no CHD, High**
$3110(0.97758) = 3040.27$	$4731(0.97758) = 4624.92$	$633(0.97758) = 618.81$

The 10% and Large Counts conditions for the chi-square test for independence are the same as for the homogeneity test. There is a slight difference in the Random condition for the two tests: a test for independence uses data from one sample but a test for homogeneity uses data from two or more samples/groups.

CONDITIONS FOR PERFORMING A CHI-SQUARE TEST FOR INDEPENDENCE

- **Random:** The data come from a well-designed random sample or randomized experiment.
  - **10%:** When sampling without replacement, check that $n \leq \frac{1}{10}N$.
- **Large Counts:** All expected counts are at least 5.

If the Random, 10%, and Large Counts conditions are met, the χ^2 statistic calculated from a two-way table can be used to perform a test of H_0: There is no association between two categorical variables in the population of interest. *P*-values for this test come from a chi-square distribution with df = (number of rows − 1) × (number of columns − 1). This new procedure is known as a chi-square test for independence.

CHI-SQUARE TEST FOR INDEPENDENCE

The chi-square test for independence is also known as the chi-square test for association.

Suppose the conditions are met. You can use the **chi-square test for independence** to test

H_0: There is no association between two categorical variables in the population of interest.

H_a: There is an association between two categorical variables in the population of interest.

Or, alternatively,

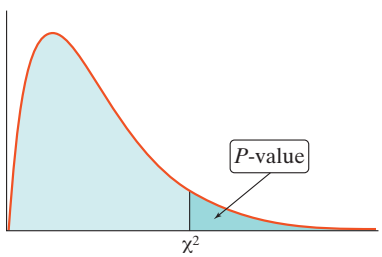
H_0: Two categorical variables are independent in the population of interest.

H_a: Two categorical variables are not independent in the population of interest.

Start by finding the expected counts. Then calculate the chi-square statistic

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

where the sum is over all cells in the two-way table. If H_0 is true, the χ^2 statistic has approximately a chi-square distribution with degrees of freedom = (number of rows − 1)(number of columns − 1). The *P*-value is the area to the right of χ^2 under the corresponding chi-square density curve.

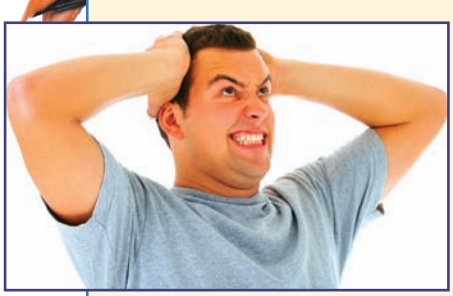


Now we're ready to complete the significance test for the anger and heart disease study.

EXAMPLE Angry People and Heart Disease

STEP 4

Chi-square test for independence



Here is the complete table of observed and expected counts for the CHD and anger study side by side:

	Observed			Expected		
	Low	Moderate	High	Low	Moderate	High
CHD	53	110	27	69.73	106.08	14.19
No CHD	3057	4621	606	3040.27	4624.92	618.81

Do the data provide convincing evidence of an association between anger level and heart disease in the population of interest?

STATE: We want to perform a test of

H_0: There is no association between anger level and heart-disease status in the population of people with normal blood pressure.
H_a: There is an association between anger level and heart-disease status in the population of people with normal blood pressure.

Because no significance level was stated, we'll use $\alpha = 0.05$.

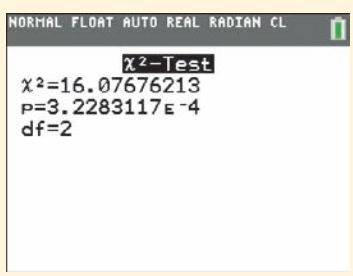
PLAN: If conditions are met, we should carry out a chi-square test for independence.

- *Random:* The data came from a random sample of 8474 people with normal blood pressure.
  - 10%: Because the researchers sampled without replacement, we need to check that the total number of people in the population with normal blood pressure is at least $10(8474) = 84{,}740$. This seems reasonable to assume.
- *Large Counts:* All the expected counts are at least 5 (the smallest is 14.19), so this condition is met.

DO: We perform calculations assuming H_0 is true.

- *Test statistic:*

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$
$$= \frac{(53 - 69.73)^2}{69.73} + \frac{(110 - 106.08)^2}{106.08} + \cdots + \frac{(606 - 618.81)^2}{618.81}$$
$$= 4.014 + 0.145 + \cdots + 0.265 = 16.077$$



- *P-value:* The two-way table of anger level versus heart disease has 2 rows and 3 columns. We will use the chi-square distribution with $df = (2 - 1)(3 - 1) = 2$ to find the P-value. Look at the $df = 2$ line in Table C. The observed statistic $\chi^2 = 16.077$ is larger than the critical value 15.20 for $\alpha = 0.0005$. So the P-value is less than 0.0005.

Using Technology: The calculator's χ^2-Test gives $\chi^2 = 16.077$ and P-value = 0.00032 using $df = 2$.

CONCLUDE: Because the P-value of 0.00032 is less than $\alpha = 0.05$, we reject H_0. We have convincing evidence of an association between anger level and heart-disease status in the population of people with normal blood pressure.

For Practice *Try Exercise* 45

A follow-up analysis reveals that two cells contribute most of the chi-square statistic: Low anger, CHD (4.014) and High anger, CHD (11.564). A much smaller number of low-anger people developed CHD than expected. And a much larger number of high-anger people got CHD than expected.

Can we conclude that proneness to anger *causes* heart disease? No. The anger and heart-disease study is an observational study, not an experiment. It isn't surprising that some other variables are confounded with anger level. For example, people prone to anger are more likely than others to be men who drink and smoke. We don't know whether the increased rate of heart disease among those with higher anger levels in the study is due to their anger or perhaps to their drinking and smoking or maybe even to gender.

CHECK YOUR UNDERSTANDING

Many popular businesses are franchises—think of McDonald's. The owner of a local franchise benefits from the brand recognition, national advertising, and detailed guidelines provided by the franchise chain. In return, he or she pays fees to the franchise firm and agrees to follow its policies. The relationship between the local owner and the franchise firm is spelled out in a detailed contract.

One clause that the contract may or may not contain is the entrepreneur's right to an exclusive territory. This means that the new outlet will be the only representative of the franchise in a specified territory and will not have to compete with other outlets of the same chain. How does the presence of an exclusive-territory clause in the contract relate to the survival of the business?

A study designed to address this question collected data from a random sample of 170 new franchise firms. Two categorical variables were measured for each franchisor. First, the franchisor was classified as successful or not based on whether or not it was still offering franchises as of a certain date. Second, the contract each franchisor offered to franchisees was classified according to whether or not there was an exclusive-territory clause. Here are the count data, arranged in a two-way table:[9]

	Exclusive Territory		
Success	**Yes**	**No**	**Total**
Yes	108	15	**123**
No	34	13	**47**
Total	**142**	**28**	**170**

Do these data provide convincing evidence at the $\alpha = 0.01$ level of an association between an exclusive-territory clause and business survival for new franchise firms?

Using Chi-Square Tests Wisely

AP® EXAM TIP If you have trouble distinguishing the two types of chi-square tests for two-way tables, you're better off just saying "chi-square test" than choosing the wrong type. Better yet, learn to tell the difference!

Both the chi-square test for homogeneity and the chi-square test for independence start with a two-way table of observed counts. They even calculate the test statistic, degrees of freedom, and *P*-value in the same way. *The questions that these two tests answer are different, however.* A chi-square test for homogeneity tests whether the distribution of a categorical variable is the same for each of several populations or treatments. The chi-square test for independence tests whether two categorical variables are associated in some population of interest.

Unfortunately, it is quite common to see questions asking about association when a test for homogeneity applies and questions asking about differences between proportions or the distribution of a variable when a test of independence applies. Sometimes, people avoid the distinction altogether and pose questions about the "relationship" between two variables.

Instead of focusing on the question asked, it's much easier to look at how the data were produced. If the data come from two or more independent random samples or treatment groups in a randomized experiment, then do a chi-square test for homogeneity. If the data come from a single random sample, with the individuals classified according to two categorical variables, use a chi-square test for independence.

EXAMPLE Scary Movies and Fear

Choosing the right type of chi-square test

Are men and women equally likely to suffer lingering fear from watching scary movies as children? Researchers asked a random sample of 117 college students to write narrative accounts of their exposure to scary movies before the age of 13. More than one-fourth of the students said that some of the fright symptoms are still present when they are awake.[10] The following table breaks down these results by gender.

	Gender		
Fright symptoms?	Male	Female	**Total**
Yes	7	29	**36**
No	31	50	**81**
Total	**38**	**79**	**117**

Minitab output for a chi-square test using these data is shown below.

Chi-Square Test: Male, Female

Expected counts are printed below observed counts

Chi-Square contributions are printed below expected counts

	Male	Female	Total
1	7	29	36
	11.69	24.31	
	1.883	0.906	
2	31	50	81
	26.31	54.69	
	0.837	0.403	
Total	38	79	117

Chi-Sq = 4.028, DF = 1, P-Value = 0.045

PROBLEM: Assume that the conditions for performing inference are met.

(a) Explain why a chi-square test for independence and not a chi-square test for homogeneity should be used in this setting.

(b) State an appropriate pair of hypotheses for researchers to test in this setting.

(c) Which cell contributes most to the chi-square statistic? In what way does this cell differ from what the null hypothesis suggests?

(d) Interpret the *P*-value in context. What conclusion would you draw at $\alpha = 0.01$?

SOLUTION:

(a) The data were produced using a single random sample of college students, who were then classified by gender and whether or not they had lingering fright symptoms. The chi-square test for homogeneity requires independent random samples from each population.

(b) The null hypothesis is H_0: There is no association between gender and ongoing fright symptoms in the population of college students. The alternative hypothesis is H_a: There is an association between gender and ongoing fright symptoms in the population of college students.

(c) Men who admit to having lingering fright symptoms account for the largest component of the chi-square statistic: 1.883 of the total 4.028. Far fewer men in the sample admitted to fright symptoms (7) than we would expect if H_0 were true (11.69).

(d) If gender and ongoing fright symptoms really are independent in the population of interest, there is a 0.045 chance of obtaining a random sample of 117 students that gives a chi-square statistic of 4.028 or higher. Because the P-value, 0.045, is greater than 0.01, we would fail to reject H_0. We do not have convincing evidence that there is an association between gender and fright symptoms in the population of college students.

For Practice *Try Exercise* 47

What if we want to compare two proportions? Shopping at secondhand stores is becoming more popular and has even attracted the attention of business schools. A study of customers' attitudes toward secondhand stores interviewed separate random samples of shoppers at two secondhand stores of the same chain in two cities. The two-way table shows the breakdown of respondents by gender.[11]

	City 1	City 2
Men	38	68
Women	203	150
Total	**241**	**218**

Do the data provide convincing evidence of a difference in the true gender distributions of shoppers at the two stores?

To answer this question, we could perform a chi-square test for homogeneity. Our hypotheses are

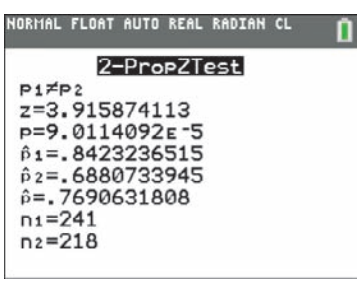
H_0: There is no difference in the true gender distributions of shoppers at the two stores.

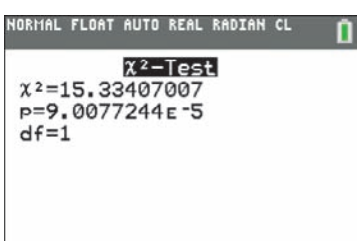
H_a: There is a difference in the true gender distributions of shoppers at the two stores.

But a difference in gender distributions would mean that there is a difference in the true proportions of female shoppers at the two stores. So we could also use a two-sample z test from Section 10.1 to compare two proportions. The hypotheses for this test are

$$H_0: p_1 - p_2 = 0$$
$$H_a: p_1 - p_2 \neq 0$$

where p_1 and p_2 are the true proportions of women shoppers at Store 1 and Store 2, respectively.





The TI-84 screen shots in the margin show the results from a two-sample z test for $p_1 - p_2$ and from a chi-square test for homogeneity. (We checked that the Random, 10%, and Large Counts conditions are met before carrying out the calculations.)

Note that the P-values from the two tests are the same except for rounding errors. You can also check that the chi-square statistic is the square of the two-sample z statistic: $(3.915\ldots)^2 = 15.334$.

As the previous example suggests, the chi-square test for homogeneity based on a 2×2 two-way table is equivalent to the two-sample z test for $p_1 - p_2$ with a two-sided alternative hypothesis. *We cannot use a chi-square test for a one-sided alternative hypothesis.* The two-sample z procedures allow us

to perform one-sided tests and to construct confidence intervals for the difference between proportions. For that reason, we recommend the Chapter 10 methods for comparing two proportions whenever you are given a choice.

Grouping quantitative data into categories As we mentioned in Chapter 1, it is possible to convert a quantitative variable to a categorical variable by grouping together intervals of values. Here's an example. Researchers surveyed independent random samples of shoppers at two secondhand stores of the same chain in two cities. The two-way table below summarizes data on the incomes of the shoppers in the two samples.

Income	City 1	City 2
Under $10,000	70	62
$10,000 to $19,999	52	63
$20,000 to $24,999	69	50
$25,000 to $34,999	22	19
$35,000 or more	28	24

Personal income is a quantitative variable. But by grouping the values of this variable, we create a categorical variable. We could use these data to carry out a chi-square test for homogeneity because the data came from independent random samples of shoppers at the two stores. Comparing the distributions of income for shoppers at the two stores would give more information than simply comparing their mean incomes.

What can we do if the expected cell counts aren't all at least 5? Let's look at a situation where this is the case. A sample survey asked a random sample of young adults, "Where do you live now? That is, where do you stay most often?" A two-way table of all 2984 people in the sample (both men and women) classified by their age and by where they lived is shown below.[12] Living arrangement is a categorical variable. Even though age is quantitative, the two-way table treats age as dividing the young adults into four categories. The table gives the observed counts for all 20 combinations of age and living arrangement.

	Age (years)				
Living arrangement	**19**	**20**	**21**	**22**	**Total**
Parents' home	324	378	337	318	**1357**
Another person's home	37	47	40	38	**162**
Your own place	116	279	372	487	**1254**
Group quarters	58	60	49	25	**192**
Other	5	2	3	9	**19**
Total	**540**	**766**	**801**	**877**	**2984**

Our null hypothesis is H_0: There is no association between age and living arrangement in the population of young adults. The table below shows the expected counts assuming H_0 is true. We can see that two of the expected counts (circled in red) are less than 5. This violates the Large Counts condition.

	Age (years)				
Living arrangement	**19**	**20**	**21**	**22**	**Total**
Parents' home	245.57	348.35	364.26	398.82	**1357**
Another person's home	29.32	41.59	43.49	47.61	**162**
Your own place	226.93	321.90	336.61	368.55	**1254**
Group quarters	34.75	49.29	51.54	56.43	**192**
Other	3.44	4.88	5.10	5.58	**19**
Total	**540**	**766**	**801**	**877**	**2984**

A clever strategy is to "collapse" the table by combining two or more rows or columns. In this case, it might make sense to combine the Group quarters and Other living arrangements. Doing so and then running a chi-square test in Minitab gives the following output. Notice that the Large Counts condition is now met.

```
Chi-Square Test: 19, 20, 21, 22

Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

             19          20          21          22       Total
1           324         378         337         318        1357
         245.57      348.35      364.26      398.82
         25.049       2.525       2.040      16.379

2            37          47          40          38         162
          29.32       41.59       43.49       47.61
          2.014       0.705       0.279       1.940

3           116         279         372         487        1254
         226.93      321.90      336.61      368.55
         54.226       5.719       3.720      38.068

4            63          62          52          34         211
          38.18       54.16       56.64       62.01
         16.129       1.134       0.380      12.654

Total       540         766         801         877        2984

Chi-Sq = 182.961, DF = 9, P-Value = 0.000
```

case closed

Do Dogs Resemble Their Owners?

In the chapter-opening Case Study (page 677), we described a study that investigated whether or not dogs resemble their owners. The researchers who conducted the experiment believe that resemblance between dog and owner might differ for purebred and mixed-breed dogs. Here is a two-way table summarizing the results of the experiment:

	Breed status	
Resemblance?	**Purebred dogs**	**Mixed-breed dogs**
Resemble owner	16	7
Don't resemble	9	13

1. Why did researchers photograph a random sample of dogs and their owners in this study?

Do the data from this study provide convincing evidence of an association between dogs' breed status and whether or not they resemble their owners? Questions 2 through 5 address this issue.

2. Which type of chi-square test should be used to help answer the question of interest? State an appropriate pair of hypotheses for the test you choose.
3. The table shows the expected counts for the appropriate chi-square test in Question 2.

	Breed status	
Resemblance?	**Purebred dogs**	**Mixed-breed dogs**
Resemble owner	12.78	10.22
Don't resemble	12.22	9.78

   (a) Show how the expected count for the cell "purebred dogs, resemble owner" was computed.
   (b) Explain why the Large Counts condition is met.

4. Find the test statistic and *P*-value. Be sure to state the degrees of freedom you are using.
5. What conclusion would you draw?

Section 11.2 Summary

- We can use a two-way table to summarize data involving two categorical variables. To analyze the data, we compare the conditional distributions of one variable for each value of the other variable. Then we turn to formal inference. Two different ways of producing data for two-way tables lead to two different types of chi-square tests.
- Some studies aim to compare the distribution of a single categorical variable for each of several populations or treatments. In such cases, researchers should take independent random samples from the populations of interest or use the groups in a randomized experiment. The null hypothesis is that there is no difference in the distribution of the categorical variable for each of the populations or treatments. We use the **chi-square test for homogeneity** to test this hypothesis.

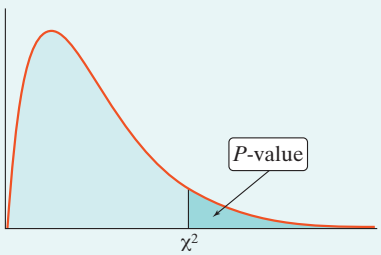
- The conditions for performing a chi-square test for homogeneity are:
  - **Random:** The data come from independent random samples or the groups in a randomized experiment.
    - **10%:** When sampling without replacement, check that the population is at least 10 times as large as the sample.
  - **Large Counts:** All expected counts must be at least 5.
- Other studies are designed to investigate the relationship between two categorical variables. In such cases, researchers take a random sample from the population of interest and classify each individual based on the two categorical variables. The **chi-square test for independence** tests the null hypothesis that there is no association between the two categorical variables in the population of interest. Another way to state the null hypothesis is H_0: The two categorical variables are independent in the population of interest.
- The conditions for performing a chi-square test for independence are:
  - **Random:** The data come from a well-designed random sample or randomized experiment.
    - **10%:** When sampling without replacement, check that the population is at least 10 times as large as the sample.
  - **Large Counts:** All expected counts must be at least 5.
- The **expected count** in any cell of a two-way table when H_0 is true is

$$\text{expected count} = \frac{\text{row total} \cdot \text{column total}}{\text{table total}}$$

- The **chi-square statistic** is

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

where the sum is over all cells in the two-way table.



- Both types of chi-square tests for two-way tables compare the value of the statistic χ^2 with critical values from the chi-square distribution with df = (number of rows − 1)(number of columns − 1). Large values of χ^2 are evidence against H_0 and in favor of H_a, so the P-value is the area under the chi-square density curve to the right of χ^2.
- If the test finds a statistically significant result, consider doing a follow-up analysis that compares the observed and expected counts and that looks for the largest components of the chi-square statistic.

11.2 TECHNOLOGY CORNER

TI-Nspire Instructions in Appendix B; HP Prime instructions on the book's Web site.

Section 11.2 Exercises

27. **Why men and women play sports** Do men and women participate in sports for the same reasons? One goal for sports participants is social comparison—the desire to win or to do better than other people. Another is mastery—the desire to improve one's skills or to try one's best. A study on why students participate in sports collected data from independent random samples of 67 male and 67 female undergraduates at a large university.[13] Each student was classified into one of four categories based on his or her responses to a questionnaire about sports goals. The four categories were high social comparison–high mastery (HSC-HM), high social comparison–low mastery (HSC-LM), low social comparison–high mastery (LSC-HM), and low social comparison–low mastery (LSC-LM). One purpose of the study was to compare the goals of male and female students. Here are the data displayed in a two-way table:

pg 697

	Gender	
Goal	Female	Male
HSC-HM	14	31
HSC-LM	7	18
LSC-HM	21	5
LSC-LM	25	13

(a) Calculate the conditional distribution (in proportions) of the reported sports goals for each gender.

(b) Make an appropriate graph for comparing the conditional distributions in part (a).

(c) Write a few sentences comparing the distributions of sports goals for male and female undergraduates.

28. **How are schools doing?** The nonprofit group Public Agenda conducted telephone interviews with three randomly selected groups of parents of high school children. There were 202 black parents, 202 Hispanic parents, and 201 white parents. One question asked was "Are the high schools in your state doing an excellent, good, fair, or poor job, or don't you know enough to say?" Here are the survey results:[14]

	Black parents	Hispanic parents	White parents
Excellent	12	34	22
Good	69	55	81
Fair	75	61	60
Poor	24	24	24
Don't know	22	28	14
Total	**202**	**202**	**201**

(a) Calculate the conditional distribution (in proportions) of responses for each group of parents.

(b) Make an appropriate graph for comparing the conditional distributions in part (a).

(c) Write a few sentences comparing the distributions of responses for the three groups of parents.

29. **Why women and men play sports** Refer to Exercise 27. Do the data provide convincing evidence of a difference in the distributions of sports goals for male and female undergraduates at the university?

pg 704

(a) State appropriate null and alternative hypotheses for a significance test to help answer this question.

(b) Calculate the expected counts. Show your work.

(c) Calculate the chi-square statistic. Show your work.

30. **How are schools doing?** Refer to Exercise 28. Do the data provide convincing evidence of a difference in the distributions of opinions about how high schools are doing among black, Hispanic, and white parents?

(a) State appropriate null and alternative hypotheses for a significance test to help answer this question.

(b) Calculate the expected counts. Show your work.

(c) Calculate the chi-square statistic. Show your work.

31. **Why women and men play sports** Refer to Exercises 27 and 29.

pg 704

(a) Check that the conditions for performing the chi-square test are met.

(b) Use Table C to find the *P*-value. Then use your calculator's χ^2`cdf` command.

(c) Interpret the *P*-value from the calculator in context.

(d) What conclusion would you draw? Justify your answer.

32. **How are schools doing?** Refer to Exercises 28 and 30.

(a) Check that the conditions for performing the chi-square test are met.

(b) Use Table C to find the *P*-value. Then use your calculator's χ^2`cdf` command.

(c) Interpret the *P*-value from the calculator in context.

(d) What conclusion would you draw? Justify your answer.

33. **Python eggs** How is the hatching of water python eggs influenced by the temperature of the snake's nest? Researchers randomly assigned newly laid eggs to one of three water temperatures: hot, neutral, or cold. Hot duplicates the extra warmth provided by the mother python, and cold duplicates the absence of the mother. Here are the data on the number of eggs that hatched and didn't hatch:[15]

pg 708

	Water Temperature		
Hatched?	Cold	Neutral	Hot
Yes	16	38	75
No	11	18	29

(a) Compare the distributions of hatching status for the three treatments.

(b) Are the differences between the three groups statistically significant? Give appropriate evidence to support your answer.

34. **Don't do drugs!** Cocaine addicts need cocaine to feel any pleasure, so perhaps giving them an antidepressant drug will help. A three-year study with 72 chronic cocaine users compared an antidepressant drug called desipramine with lithium (a standard drug to treat cocaine addiction) and a placebo. One-third of the subjects were randomly assigned to receive each treatment. Here are the results:[16]

	Drug administered		
Relapsed?	Desipramine	Lithium	Placebo
Yes	10	18	20
No	14	6	4

(a) Compare the distributions of relapse status for the three treatments.

(b) Are the differences among the three groups statistically significant? Give appropriate evidence to support your answer.

35. **Sorry, no chi-square** How do U.S. residents who travel overseas for leisure differ from those who travel for business? The following is the breakdown by occupation:[17]

Occupation	Leisure travelers (%)	Business travelers (%)
Professional/technical	36	39
Manager/executive	23	48
Retired	14	3
Student	7	3
Other	20	7
Total	**100**	**100**

Explain why we can't use a chi-square test to learn whether these two distributions differ significantly.

36. **Going Nuts** The UR Nuts Company sells Deluxe and Premium nut mixes, both of which contain only cashews, brazil nuts, almonds, and peanuts. The Premium nuts are much more expensive than the Deluxe nuts. A consumer group suspects that the two nut mixes are really the same. To find out, the group took separate random samples of 20 pounds of each nut mix and recorded the weights of each type of nut in the sample. Here are the data:[18]

	Type of mix	
Type of nut	Premium	Deluxe
Cashew	6 lb	5 lb
Brazil nut	3 lb	4 lb
Almond	5 lb	6 lb
Peanut	6 lb	5 lb

Explain why we can't use a chi-square test to determine whether these two distributions differ significantly.

37. **How to quit smoking** It's hard for smokers to quit. Perhaps prescribing a drug to fight depression will work as well as the usual nicotine patch. Perhaps combining the patch and the drug will work better than either treatment alone. Here are data from a randomized, double-blind trial that compared four treatments.[19] A "success" means that the subject did not smoke for a year following the beginning of the study.

Group	Treatment	Subjects	Successes
1	Nicotine patch	244	40
2	Drug	244	74
3	Patch plus drug	245	87
4	Placebo	160	25

(a) Summarize these data in a two-way table. Then compare the success rates for the four treatments.

(b) Explain in words what the null hypothesis H_0: $p_1 = p_2 = p_3 = p_4$ says about subjects' smoking habits.

(c) Do the data provide convincing evidence of a difference in the effectiveness of the four treatments at the $\alpha = 0.05$ significance level?

38. **Preventing strokes** Aspirin prevents blood from clotting and so helps prevent strokes. The Second European Stroke Prevention Study asked whether adding another anticlotting drug named dipyridamole would be more effective for patients who had already had a stroke. Here are the data on strokes during the two years of the study:[20]

Group	Treatment	Number of patients	Number of strokes
1	Placebo	1649	250
2	Aspirin	1649	206
3	Dipyridamole	1654	211
4	Both	1650	157

(a) Summarize these data in a two-way table. Then compare the stroke rates for the four treatments.

(b) Explain in words what the null hypothesis $H_0: p_1 = p_2 = p_3 = p_4$ says about the incidence of strokes.

(c) Do the data provide convincing evidence of a difference in the effectiveness of the four treatments at the $\alpha = 0.05$ significance level?

39. **How to quit smoking** Perform a follow-up analysis of the test in Exercise 37 by finding the individual components of the chi-square statistic. Which cell(s) contributed most to the final result and in what direction?

40. **Preventing strokes** Perform a follow-up analysis of the test in Exercise 38 by finding the individual components of the chi-square statistic. Which cell(s) contributed most to the final result and in what direction?

41. **Attitudes toward recycled products** Some people believe recycled products are lower in quality than other products, a fact that makes recycling less practical. Here are data on attitudes toward coffee filters made of recycled paper from a random sample of adults:[21]

pg 711

	Recycled coffee filter status	
Quality rating	Buyers	Nonbuyers
Higher	20	29
Same	7	25
Lower	9	43

Make a well-labeled bar graph that compares buyers' and nonbuyers' opinions about recycled filters. Describe what you see.

42. **Is astrology scientific?** The General Social Survey asked a random sample of adults their opinion about whether astrology is very scientific, sort of scientific, or not at all scientific. Here is a two-way table of counts for people in the sample who had three levels of higher education:[22]

	Degree Held		
	Associate's	Bachelor's	Master's
Not at all scientific	169	256	114
Very or sort of scientific	65	65	18

Make a well-labeled bar graph that compares opinions about astrology for the three education categories. Describe what you see.

43. **Attitudes toward recycled products** Refer to Exercise 41.

(a) State appropriate hypotheses for performing a chi-square test of independence in this setting.

(b) Compute the expected counts assuming that H_0 is true. Show your work.

(c) Calculate the chi-square statistic, df, and P-value.

(d) What conclusion would you draw?

44. **Is astrology scientific?** Refer to Exercise 42.

(a) State appropriate hypotheses for performing a chi-square test of independence in this setting.

(b) Compute the expected counts assuming that H_0 is true. Show your work.

(c) Calculate the chi-square statistic, df, and P-value.

(d) What conclusion would you draw?

45. **Regulating guns** The National Gun Policy Survey asked a random sample of adults, "Do you think there should be a law that would ban possession of handguns except for the police and other authorized persons?" Here are the responses, broken down by the respondent's level of education:[23]

pg 715

	Education				
	Less than high school	High school grad	Some college	College grad	Postgrad degree
Yes	58	84	169	98	77
No	58	129	294	135	99

Does the sample provide convincing evidence of an association between education level and opinion about a handgun ban in the adult population?

46. **Market research** Before bringing a new product to market, firms carry out extensive studies to learn how consumers react to the product and how best to advertise its advantages. Here are data from a study of a new laundry detergent.[24] The participants are a random sample of people who don't currently use the established brand that the new product will compete with. Give subjects free samples of both detergents. After they have tried both for a while, ask which they prefer. The answers may depend on other facts about how people do laundry.

	Laundry Practices			
	Soft water, warm wash	Soft water, hot wash	Hard water, warm wash	Hard water, hot wash
Prefer standard product	53	27	42	30
Prefer new product	63	29	68	42

Does the sample provide convincing evidence of an association between laundry practices and product preference in the population of interest?

47. **Where do young adults live?** A survey by the pg 718 National Institutes of Health asked a random sample of young adults (aged 19 to 25 years), "Where do you live now? That is, where do you stay most often?" Here is the full two-way table (omitting a few who refused to answer and one who claimed to be homeless):[25]

	Female	Male
Parents' home	923	986
Another person's home	144	132
Own place	1294	1129
Group quarters	127	119

(a) Should we use a chi-square test for homogeneity or a chi-square test for independence in this setting? Justify your answer.

(b) State appropriate hypotheses for performing the type of test you chose in part (a).

Minitab output from a chi-square test is shown below.

```
Chi-Square Test: Female, Male
Expected counts are printed below observed
counts
Chi-Square contributions are printed below
expected counts
             Female       Male    Total
   1            923        986     1909
             978.49     930.51
              3.147      3.309

   2            144        132      276
             141.47     134.53
              0.045      0.048

   3           1294       1129     2423
            1241.95    1181.05
              2.181      2.294

   4            127        119      246
             126.09     119.91
              0.007      0.007

   Total       2488       2366     4854
Chi-Sq = 11.038, DF = 3, P-Value = 0.012
```

(c) Check that the conditions for carrying out the test are met.

(d) Interpret the *P*-value in context. What conclusion would you draw?

48. **Students and catalog shopping** What is the most important reason that students buy from catalogs? The answer may differ for different groups of students. Here are results for separate random samples of American and Asian students at a large midwestern university:[26]

	American	Asian
Save time	29	10
Easy	28	11
Low price	17	34
Live far from stores	11	4
No pressure to buy	10	3

(a) Should we use a chi-square test for homogeneity or a chi-square test for independence in this setting? Justify your answer.

(b) State appropriate hypotheses for performing the type of test you chose in part (a).

Minitab output from a chi-square test is shown below.

```
Chi-Square Test: American, Asian
Expected counts are printed below observed
counts
Chi-Square contributions are printed below
expected counts
             American      Asian    Total
   1               29         10       39
                23.60      15.40
                1.236      1.894

   2               28         11       39
                23.60      15.40
                0.821      1.258

   3               17         34       51
                30.86      20.14
                6.225      9.538

   4               11          4       15
                 9.08       5.92
                0.408      0.625

   5               10          3       13
                 7.87       5.13
                0.579      0.887

   Total           95         62      157
Chi-Sq = 23.470, DF = 4, P-Value = 0.0001
```

(c) Check that the conditions for carrying out the test are met.

(d) Interpret the *P*-value in context. What conclusion would you draw?

49. **Treating ulcers** Gastric freezing was once a recommended treatment for ulcers in the upper intestine. Use of gastric freezing stopped after experiments showed it had no effect. One randomized comparative experiment found that 28 of the 82 gastric-freezing patients improved, while 30 of

the 78 patients in the placebo group improved.[27] We can test the hypothesis of "no difference" in the effectiveness of the treatments in two ways: with a two-sample z test or with a chi-square test.

(a) Minitab output for a chi-square test is shown below. State appropriate hypotheses and interpret the P-value in context. What conclusion would you draw?

```
Chi-Square Test: Gastric freezing, Placebo
Expected counts are printed below observed
counts
Chi-Square contributions are printed below
expected counts
     Gastric freezing    Placebo    Total
1                  28         30       58
                29.73      28.27
                0.100      0.105
2                  54         48      102
                52.27      49.73
                0.057      0.060
Total              82         78      160
Chi-Sq = 0.322, DF = 1, P-Value = 0.570
```

(b) Minitab output for a two-sample z test is shown below. Explain how these results are consistent with the test in part (a).

```
Test for Two Proportions
Sample      X       N       Sample p
1          28      82       0.341463
2          30      78       0.384615
Difference = p (1) − p (2)
Estimate for difference: −0.0431520
Test for difference = 0 (vs not = 0):
Z = −0.57 P-Value = 0.570
```

50. Opinions about the death penalty The General Social Survey asked separate random samples of people with only a high school degree and people with a bachelor's degree, "Do you favor or oppose the death penalty for persons convicted of murder?" The following table gives the responses of people whose highest education was a high school degree and of people with a bachelor's degree:

	Highest education level	
	High school	Bachelor's degree
Favor	1010	319
Oppose	369	185

We can test the hypothesis of "no difference" in support for the death penalty among people in these educational categories in two ways: with a two-sample z test or with a chi-square test.

(a) Minitab output for a chi-square test is shown below. State appropriate hypotheses and interpret the P-value in context. What conclusion would you draw?

```
Chi-Square Test: C1, C2
Expected counts are printed below
observed counts
Chi-Square contributions are printed be-
low expected counts
            C1        C2    Total
1         1010       319     1329
        973.28    355.72
         1.385     3.790
2          369       185      554
        405.72    148.28
         3.323     9.092
Total     1379       504     1883
Chi-Sq = 17.590, DF = 1, P-Value = 0.000
```

(b) Minitab output for a two-sample z test is shown below. Explain how these results are consistent with the test in part (a).

```
Test for Two Proportions
Sample      X       N     Sample p
1        1010    1379     0.732415
2         319     504     0.632937
Difference = p (1) − p (2)
Estimate for difference: 0.0994783
Test for difference = 0 (vs not = 0):
Z = 4.19 P-Value = 0.000
```

Multiple choice: Select the best answer for Exercises 51 to 56.

Exercises 51 to 55 refer to the following setting. The National Longitudinal Study of Adolescent Health interviewed a random sample of 4877 teens (grades 7 to 12). One question asked was "What do you think are the chances you will be married in the next ten years?" Here is a two-way table of the responses by gender:[28]

	Female	Male
Almost no chance	119	103
Some chance, but probably not	150	171
A 50–50 chance	447	512
A good chance	735	710
Almost certain	1174	756

51. Which of the following would be the most appropriate type of graph for these data?

(a) A bar chart showing the marginal distribution of opinion about marriage

(b) A bar chart showing the marginal distribution of gender

(c) A bar chart showing the conditional distribution of gender for each opinion about marriage

(d) A bar chart showing the conditional distribution of opinion about marriage for each gender

(e) Dotplots that display the number in each opinion category for each gender

52. The appropriate null hypothesis for performing a chi-square test is that

(a) equal proportions of female and male teenagers are almost certain they will be married in 10 years.

(b) there is no difference between the distributions of female and male teenagers' opinions about marriage in this sample.

(c) there is no difference between the distributions of female and male teenagers' opinions about marriage in the population.

(d) there is no association between gender and opinion about marriage in the sample.

(e) there is no association between gender and opinion about marriage in the population.

53. The expected count of females who respond "almost certain" is

(a) 487.7. **(c)** 965. **(e)** 1174.

(b) 525. **(d)** 1038.8.

54. The degrees of freedom for the chi-square test for this two-way table are

(a) 4. **(c)** 10. **(e)** 4876.

(b) 8. **(d)** 20.

55. For these data, $\chi^2 = 69.8$ with a P-value of approximately 0. Assuming that the researchers used a significance level of 0.05, which of the following is true?

(a) A Type I error is possible.

(b) A Type II error is possible.

(c) Both a Type I and a Type II error are possible.

(d) There is no chance of making a Type I or Type II error because the P-value is approximately 0.

(e) There is no chance of making a Type I or Type II error because the calculations are correct.

56. When analyzing survey results from a two-way table, the main distinction between a test for independence and a test for homogeneity is

(a) how the degrees of freedom are calculated.

(b) how the expected counts are calculated.

(c) the number of samples obtained.

(d) the number of rows in the two-way table.

(e) the number of columns in the two-way table.

For Exercises 57 and 58, you may find the inference summary chart inside the back cover helpful.

57. **Inference recap** (8.1 to 11.2) In each of the following settings, state which inference procedure from Chapter 8, 9, 10, or 11 you would use. Be specific. For example, you might say "two-sample z test for the difference between two proportions." You do not need to carry out any procedures.[29]

(a) What is the average voter turnout during an election? A random sample of 38 cities was asked to report the percent of registered voters who actually voted in the most recent election.

(b) Are blondes more likely to have a boyfriend than the rest of the single world? Independent random samples of 300 blondes and 300 nonblondes were asked whether they have a boyfriend.

58. **Inference recap** (8.1 to 11.2) In each of the following settings, state which inference procedure from Chapter 8, 9, 10, or 11 you would use. Be specific. For example, you might say "two-sample z test for the difference between two proportions." You do not need to carry out any procedures.[30]

(a) Is there a relationship between attendance at religious services and alcohol consumption? A random sample of 1000 adults was asked whether they regularly attend religious services and whether they drink alcohol daily.

(b) Separate random samples of 75 college students and 75 high school students were asked how much time, on average, they spend watching television each week. We want to estimate the difference in the average amount of TV watched by high school and college students.

Exercises 59 to 60 refer to the following setting. For their final project, a group of AP® Statistics students investigated the following question: "Will changing the rating scale on a survey affect how people answer the question?" To find out, the group took an SRS of 50 students from an alphabetical roster of the school's just over 1000 students. The first 22 students chosen were asked to rate the cafeteria food on a scale of 1 (terrible) to 5 (excellent). The remaining 28 students were asked to rate the cafeteria food on a scale of 0 (terrible) to 4 (excellent). Here are the data:

	1 to 5 scale				
Rating	1	2	3	4	5
Frequency	2	3	1	13	3

	0 to 4 scale				
Rating	0	1	2	3	4
Frequency	0	0	2	18	8

59. **Design and analysis** (4.2)

(a) Was this an observational study or an experiment? Justify your answer.

(b) Explain why it would not be appropriate to perform a chi-square test in this setting.

60. **Average ratings** (1.3, 10.2) The students decided to compare the average ratings of the cafeteria food on the two scales.

(a) Find the mean and standard deviation of the ratings for the students who were given the 1-to-5 scale.

(b) For the students who were given the 0-to-4 scale, the ratings have a mean of 3.21 and a standard deviation of 0.568. Since the scales differ by one point, the group decided to add 1 to each of these ratings. What are the mean and standard deviation of the adjusted ratings?

(c) Would it be appropriate to compare the means from parts (a) and (b) using a two-sample t test? Justify your answer.

FRAPPY! Free Response AP® Problem, Yay!

The following problem is modeled after actual AP® Statistics exam free response questions. Your task is to generate a complete, concise response in 15 minutes.

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

Two statistics students wanted to know if including additional information in a survey question would change the distribution of responses. To find out, they randomly selected 30 teenagers and asked them one of the following two questions. Fifteen of the teenagers were randomly assigned to answer Question A, and the other 15 students were assigned to answer Question B.

A: When choosing a college, how important is a good athletic program: very important, important, somewhat important, not that important, or not important at all?

B: It's sad that some people choose a college based on its athletic program. When choosing a college, how important is a good athletic program: very important, important, somewhat important, not that important, or not important at all?

The table below summarizes the responses to both questions. For these data, the chi-square test statistic is $\chi^2 = 6.12$.

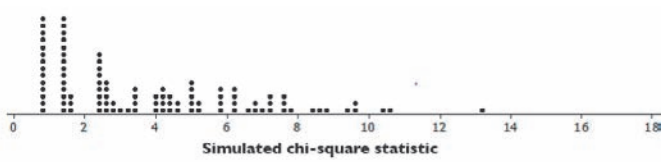
	Question A	Question B	Total
Very important	7	2	**9**
Important	4	3	**7**
Somewhat important	2	3	**5**
Not that important	1	2	**3**
Not important at all	1	5	**6**
Total	**15**	**15**	**30**

(a) State the hypotheses that the students are interested in testing.

(b) Describe a Type I error and a Type II error in the context of the hypotheses stated in part (a).

(c) For these data, explain why it would *not* be appropriate to use a chi-square distribution to calculate the P-value.

(d) To estimate the P-value, 100 trials of a simulation were conducted, assuming that the additional information didn't have an effect on the response to the question. In each trial of the simulation, the value of the chi-square statistic was calculated. These simulated chi-square statistics are displayed in the dotplot below.



Based on the results of the simulation, what conclusion would you make about the hypotheses stated in part (a)?

After you finish, you can view two example solutions on the book's Web site (www.whfreeman.com/tps5e). Determine whether you think each solution is "complete," "substantial," "developing," or "minimal." If the solution is not complete, what improvements would you suggest to the student who wrote it? Finally, your teacher will provide you with a scoring rubric. Score your response and note what, if anything, you would do differently to improve your own score.

Chapter Review

Section 11.1: Chi-Square Tests for Goodness of Fit

In this section, you learned the details for performing a chi-square test for goodness of fit. The null hypothesis is that a single categorical variable follows a specified distribution. The alternative hypothesis is that the variable does not follow the specified distribution.

The chi-square statistic measures the difference between the observed distribution of a categorical variable and its hypothesized distribution. To calculate the chi-square statistic, use the following formula that involves the observed and expected counts for each value of the categorical variable:

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

To calculate the expected counts, multiply the total sample size by the hypothesized proportion for each category. Larger values of the chi-square statistic provide more convincing evidence that the categorical variable does not have the hypothesized distribution.

When the Random, 10%, and Large Counts conditions are satisfied, we can accurately model the sampling distribution of a chi-square statistic using a chi-square distribution (density curve). The Random condition says that the data are from a well-designed random sample or a randomized experiment. The 10% condition says that the sample size should be at most 10% of the population size when sampling without replacement. The Large Counts condition says that the *expected* counts for each category must be at least 5. In a test for goodness of fit, use a chi-square distribution with degrees of freedom = number of categories − 1.

When the results of a test for goodness of fit are significant, consider doing a follow-up analysis. Identify which categories of the variable had the largest contributions to the chi-square statistic and whether the observed values in those categories were larger or smaller than expected.

Section 11.2: Inference for Two-Way Tables

In this section, you learned how to perform inference for categorical data that are summarized in a two-way table. To begin the analysis, compare the conditional distributions of the response variable for each value of the explanatory variable. Displaying these distributions with a bar graph will help you make an effective comparison.

There are two types of chi-square tests that could apply when data are summarized in a two-way table. A test for homogeneity compares the distribution of a single categorical variable for two or more populations or treatments. A test for independence looks for an association between two categorical variables in a single population.

In a chi-square test for homogeneity, the null hypothesis is that there is no difference between the true distributions of a categorical variable for two or more populations or treatments. The alternative hypothesis is that there is a difference in the distributions. The Random condition is that the data come from independent random samples or groups in a randomized experiment. The 10% condition applies when sampling without replacement, but not in experiments. Finally, the Large Counts condition remains the same—the expected counts must be at least 5 in each cell of the two-way table.

To calculate the expected counts for a test for homogeneity, use the following formula:

$$\text{expected count} = \frac{\text{row total} \cdot \text{column total}}{\text{table total}}$$

To calculate the *P*-value, compute the chi-square statistic and use a chi-square distribution with degrees of freedom = (number of rows − 1)(number of columns − 1).

In a chi-square test for independence, the null hypothesis is that there is no association between two categorical variables in one population. The alternative hypothesis is that there is an association between the two variables. For this test, the Random condition says that the data must come from a single random sample. The 10% condition applies when sampling without replacement. The Large Counts condition is still the same—the expected counts must all be at least 5. The method for calculating expected counts, the chi-square statistic, the degrees of freedom, and the *P*-value are exactly the same in a test for independence and a test for homogeneity.

As with tests for goodness of fit, when the results of a test for homogeneity or independence are significant, consider doing a follow-up analysis. Identify which cells in the two-way table had the largest contributions to the chi-square statistic and whether the observed values in those cells were larger or smaller than expected.

What Did You Learn?

Learning Objective	Section	Related Example on Page(s)	Relevant Chapter Review Exercise(s)
State appropriate hypotheses and compute expected counts for a chi-square test for goodness of fit.	11.1	681	R11.1
Calculate the chi-square statistic, degrees of freedom, and *P*-value for a chi-square test for goodness of fit.	11.1	683, 685	R11.1
Perform a chi-square test for goodness of fit.	11.1	688	R11.1
Conduct a follow-up analysis when the results of a chi-square test are statistically significant.	11.1,11.2	Discussion on 690–691	R11.4
Compare conditional distributions for data in a two-way table.	11.2	697, 711	R11.3, R11.5
State appropriate hypotheses and compute expected counts for a chi-square test based on data in a two-way table.	11.2	701, 713	R11.2, R11.3, R11.4, R11.5
Calculate the chi-square statistic, degrees of freedom, and *P*-value for a chi-square test based on data in a two-way table.	11.2	704	R11.3, R11.5
Perform a chi-square test for homogeneity.	11.2	708	R11.3
Perform a chi-square test for independence.	11.2	715	R11.5
Choose the appropriate chi-square test.	11.2	718	R11.4

Chapter 11 Chapter Review Exercises

These exercises are designed to help you review the important ideas and methods of the chapter.

R11.1 Testing a genetic model Biologists wish to cross pairs of tobacco plants having genetic makeup Gg, indicating that each plant has one dominant gene (G) and one recessive gene (g) for color. Each offspring plant will receive one gene for color from each parent. The Punnett square below shows the possible combinations of genes received by the offspring:

		Parent 2 passes on:	
		G	g
Parent 1 passes on:	G	GG	Gg
	g	Gg	gg

The Punnett square suggests that the expected ratio of green (GG) to yellow-green (Gg) to albino (gg) tobacco plants should be 1:2:1. In other words, the biologists predict that 25% of the offspring will be green, 50% will be yellow-green, and 25% will be albino. To test their hypothesis about the distribution of offspring, the biologists mate 84 randomly selected pairs of yellow-green parent plants. Of 84 offspring, 23 plants were green, 50 were yellow-green, and 11 were albino. Do the data provide convincing evidence at the $\alpha = 0.01$ level that the true distribution of offspring is different from what the biologists predict?

R11.2 Sorry, no chi-square We would prefer to learn from teachers who know their subject. Perhaps even preschool children are affected by how knowledgeable they think teachers are. Assign 48 three- and four-year-olds at random to be taught the name of a new toy by either an adult who claims to know about the toy or an adult who claims not to know about it. Then ask the children to pick out a picture of the new toy in a set of pictures of other toys and say its name. The response variable is the count of right answers in four tries. Here are the data:[31]

	Correct Answers				
	0	1	2	3	4
Knowledgeable teacher	5	1	6	3	9
Ignorant teacher	20	0	3	0	1

The researchers report that children taught by the teacher who claimed to be knowledgeable did significantly better ($\chi^2 = 20.24$, $P < 0.05$). Explain why this result isn't valid.

R11.3 Stress and heart attacks You read a newspaper article that describes a study of whether stress management can help reduce heart attacks. The 107 subjects all had reduced blood flow to the heart and so were at risk of a heart attack. They were assigned at random to three groups. The article goes on to say:

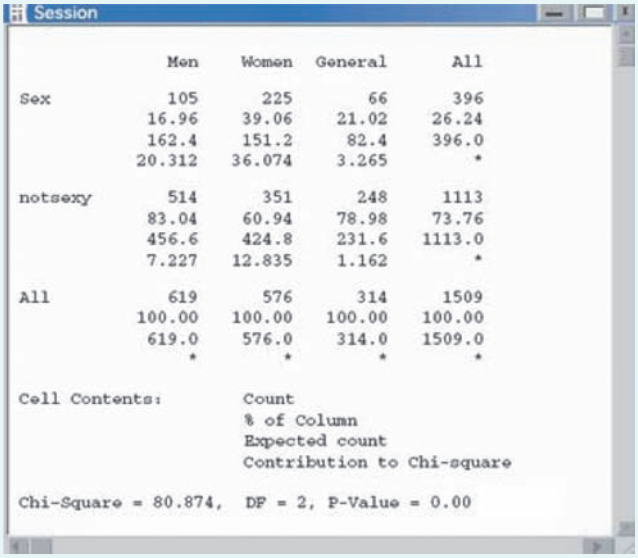
> One group took a four-month stress management program, another underwent a four-month exercise program, and the third received usual heart care from their personal physicians. In the next three years, only three of the 33 people in the stress management group suffered "cardiac events," defined as a fatal or non-fatal heart attack or a surgical procedure such as a bypass or angioplasty. In the same period, 7 of the 34 people in the exercise group and 12 out of the 40 patients in usual care suffered such events.[32]

(a) Use the information in the news article to make a two-way table that describes the study results.
(b) Compare the success rates of the three treatments in preventing cardiac events.
(c) Do the data provide convincing evidence that the true success rates are not the same for the three treatments?

R11.4 Sexy magazine ads? Researchers looked at 1509 full-page ads that show a model. The two-way table below shows the main audience of the magazines in which the ads were found (young men, young women, or young adults in general) and whether or not the ad was "sexual." This was determined based on how the model was dressed (or not dressed).[33]

	Readers		
	Men	Women	General
Sexual	105	225	66
Not sexual	514	351	248

The following figure displays Minitab output for a chi-square test using these data.



(a) Describe how these data could have been collected so that a test for homogeneity is appropriate.
(b) Describe how these data could have been collected so that a test for independence is appropriate.
(c) Show how each of the numbers 60.94, 424.8, and 12.835 was obtained for the "notsexy, Women" cell.
(d) Which cell contributes most to the chi-square statistic? How do the observed and expected counts compare for this cell?

R11.5 Popular kids Who were the popular kids at your elementary school? Did they get good grades or have good looks? Were they good at sports? A study was performed to examine the factors that determine social status for children in grades 4, 5, and 6. Researchers administered a questionnaire to a random sample of 478 students in these grades. One of the questions they asked was "What would you most like to do at school: make good grades, be good at sports, or be popular?" The two-way table below summarizes the students' responses.[34]

	Gender	
Goal	Female	Male
Grades	130	117
Popular	91	50
Sports	30	60

(a) Construct an appropriate graph to compare male and female responses. Write a few sentences describing the relationship between gender and goals.
(b) Is there convincing evidence at the $\alpha = 0.05$ level of an association between gender and goals for elementary school students?

Chapter 11 AP® Statistics Practice Test

Section I: Multiple Choice *Select the best answer for each question.*

T11.1 A chi-square test is used to test whether a 0 to 9 spinner is "fair" (that is, the outcomes are all equally likely). The spinner is spun 100 times, and the results are recorded. The degrees of freedom for the test will be

(a) 8. (b) 9. (c) 10. (d) 99. (e) None of these.

Exercises T11.2 and T11.3 refer to the following setting. Recent revenue shortfalls in a midwestern state led to a reduction in the state budget for higher education. To offset the reduction, the largest state university proposed a 25% tuition increase. It was determined that such an increase was needed simply to compensate for the lost support from the state. Separate random samples of 50 freshmen, 50 sophomores, 50 juniors, and 50 seniors from the university were asked whether they were strongly opposed to the increase, given that it was the minimum increase necessary to maintain the university's budget at current levels. Here are the results.

Strongly Opposed?	Year			
	Freshman	Sophomore	Junior	Senior
Yes	39	36	29	18
No	11	14	21	32

T11.2 Which hypotheses would be appropriate for performing a chi-square test?

(a) The null hypothesis is that the closer students get to graduation, the less likely they are to be opposed to tuition increases. The alternative is that how close students are to graduation makes no difference in their opinion.

(b) The null hypothesis is that the mean number of students who are strongly opposed is the same for each of the 4 years. The alternative is that the mean is different for at least 2 of the 4 years.

(c) The null hypothesis is that the distribution of student opinion about the proposed tuition increase is the same for each of the 4 years at this university. The alternative is that the distribution is different for at least 2 of the 4 years.

(d) The null hypothesis is that year in school and student opinion about the tuition increase in the sample are independent. The alternative is that these variables are dependent.

(e) The null hypothesis is that there is an association between year in school and opinion about the tuition increase at this university. The alternative hypothesis is that these variables are not associated.

T11.3 The conditions for carrying out the chi-square test in exercise T11.2 are

I. Independent random samples from the populations of interest.

II. All expected counts are at least 5.

III. The population sizes are at least 10 times the sample sizes.

Which of the conditions is (are) satisfied in this case?

(a) I only (c) I and II only (e) I, II, and III
(b) II only (d) II and III only

Exercises T11.4 to T11.6 refer to the following setting. A random sample of traffic tickets given to motorists in a large city is examined. The tickets are classified according to the race of the driver. The results are summarized in the following table.

Race:	White	Black	Hispanic	Other
Number of tickets:	69	52	18	9

The proportion of this city's population in each of the racial categories listed above is as follows:

Race:	White	Black	Hispanic	Other
Proportion:	0.55	0.30	0.08	0.07

We wish to test H_0: The racial distribution of traffic tickets in the city is the same as the racial distribution of the city's population.

T11.4 Assuming H_0 is true, the expected number of Hispanic drivers who would receive a ticket is

(a) 8. (b) 10.36. (c) 11. (d) 11.84. (e) 12.

T11.5 We compute the value of the χ^2 statistic to be 6.58. Assuming that the conditions for inference are met, the P-value of our test is

(a) greater than 0.20. (d) between 0.01 and 0.05.
(b) between 0.10 and 0.20. (e) less than 0.01.
(c) between 0.05 and 0.10.

T11.6 The category that contributes the largest component to the χ^2 statistic is

(a) White. (c) Hispanic.
(b) Black. (d) Other.
(e) The answer cannot be determined because this is only a sample.

Exercises T11.7 to T11.10 refer to the following setting. All current-carrying wires produce electromagnetic (EM) radiation, including the electrical wiring running into, through, and out of our homes. High-frequency EM radiation is thought to be a cause of cancer. The lower frequencies associated with household current are generally assumed to be harmless. To investigate the relationship between current configuration and type of cancer, researchers visited the addresses of a random sample of children who had died of some form of cancer (leukemia, lymphoma, or some other type) and classified the wiring configuration outside the dwelling as either a high-current configuration (HCC) or a low-current configuration (LCC). Here are the data:

	Leukemia	Lymphoma	Other cancers
HCC	52	10	17
LCC	84	21	31

Computer software was used to analyze the data. The output included the value $\chi^2 = 0.435$.

T11.7 The appropriate degrees of freedom for the χ^2 statistic is

(a) 1. (b) 2. (c) 3. (d) 4. (e) 5.

T11.8 The expected count of cases with lymphoma in homes with an HCC is

(a) $\frac{79 \cdot 31}{215}$. (b) $\frac{10 \cdot 21}{215}$. (c) $\frac{79 \cdot 31}{10}$. (d) $\frac{136 \cdot 31}{215}$.

(e) None of these.

T11.9 Which of the following may we conclude, based on the test results?

(a) There is convincing evidence of an association between wiring configuration and the chance that a child will develop some form of cancer.

(b) HCC either causes cancer directly or is a major contributing factor to the development of cancer in children.

(c) Leukemia is the most common type of cancer among children.

(d) There is not convincing evidence of an association between wiring configuration and the type of cancer that caused the deaths of children in the study.

(e) There is convincing evidence that HCC does not cause cancer in children.

T11.10 A Type I error would occur if we found convincing evidence that

(a) HCC wiring caused cancer when it actually didn't.

(b) HCC wiring didn't cause cancer when it actually did.

(c) there is no association between the type of wiring and the form of cancer when there actually is an association.

(d) there is an association between the type of wiring and the form of cancer when there actually is no association.

(e) the type of wiring and the form of cancer have a positive correlation when they actually don't.

Section II: Free Response *Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.*

T11.11 A large distributor of gasoline claims that 60% of all cars stopping at their service stations choose regular unleaded gas and that premium and supreme are each selected 20% of the time. To investigate this claim, researchers collected data from a random sample of drivers who put gas in their vehicles at the distributor's service stations in a large city. The results were as follows:

Gasoline Selected		
Regular	Premium	Supreme
261	51	88

Carry out a test of the distributor's claim at the 5% significance level.

T11.12 A study conducted in Charlotte, North Carolina, tested the effectiveness of three police responses to spouse abuse: (1) advise and possibly separate the couple, (2) issue a citation to the offender, and (3) arrest the offender. Police officers were trained to recognize eligible cases. When presented with an eligible case, a police officer called the dispatcher, who would randomly assign one of the three available treatments to be administered. There were a total of 650 cases in the study. Each case was classified according to whether the abuser was subsequently arrested within six months of the original incident.[35]

	Treatment		
Subsequent arrest?	Advise and separate	Citation	Arrest
No	187	181	175
Yes	25	43	39

(a) Explain the purpose of the random assignment in the design of this study.

(b) Construct a well-labeled graph that is suitable for comparing the effectiveness of the three treatments.

(c) State an appropriate pair of hypotheses for performing a chi-square test in this setting.

(d) Assume that all the conditions for performing the test in part (b) are met. The test yields $\chi^2 = 5.063$ and a P-value of 0.0796. Interpret this P-value in context. What conclusion should we draw from the study?

T11.13 In the United States, there is a strong relationship between education and smoking: well-educated people are less likely to smoke. Does a similar relationship hold in France? To find out, researchers recorded the level of education and smoking status of a random sample of 459 French men aged 20 to 60 years.[36] The two-way table below displays the data.

	Education		
Smoking Status	Primary School	Secondary School	University
Nonsmoker	56	37	53
Former	54	43	28
Moderate	41	27	36
Heavy	36	32	16

Is there convincing evidence of an association between smoking status and educational level among French men aged 20 to 60 years?